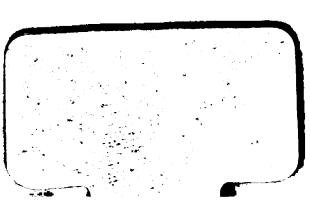

# INTRODUCTION

TO

# MATHEMATICAL STATISTICS

---

BY

CARL J. WEST, Ph. D.,

ASSISTANT PROFESSOR OF MATHEMATICS
OHIO STATE UNIVERSITY

COLUMBUS
R. G. ADAMS AND COMPANY
1918




PRESS OF
THE F. J. HEER PRINTING CO.
COLUMBUS, OHIO

## PREFACE.

IT is the aim of this book to present certain topics of elementary statistical theory which have been found useful and workable.

The statement would seem warranted that no more than the very simplest methods should be used by one who has no knowledge of the principles underlying the methods. Busy though the scientist may be, he owes it to the science and to the persons who may accept his results to have some familiarity with his tools. The blind application of formulas in statistics has been made possible by the convenient manuals that have appeared and has been encouraged by the fact that the theory has been so surrounded by intricate and involved mathematics that it was only by an extended research that a knowledge of the theory could be obtained.

There is no real reason why the theory of statistical methods should remain in obscurity. The necessary mathematics is largely elementary arithmetic and except in a few cases there is no need for higher mathematics. This book presupposes a reasonable familiarity with elementary mathematics only.

Because of the desire to eliminate higher mathematics from the body of the book the discussion of the theory of the Generalized Frequency Curves of Pearson has been deferred to Appendix I. For the same reason a discussion of the promising method of variate differences is omitted, as is the mathematical theory of random selection.

While it is hoped that the statistical data of this book may be of interest in themselves they have been selected solely with reference to their usefulness in illustrating the theory. For this reason all examples and exercises have to do with very simple data. The author will appreciate notice of such numerical and other inaccuracies as may be found.

The idea is emphasized that a formula or method to be of practical and trustworthy value to a statistician must be so simple and direct that the final results can be interpreted in terms

of the original conditions or the given data. To illustrate, if the arithmetic mean is ten per cent. larger in one distribution than in another what difference does this variation indicate in the forms of the distributions or in the values of the two series of measurements? If one correlation ratio is 0.54 and a second 0.59 how much more closely related are the attributes in the second than in the first? It must always be remembered that mathematics is but a tool to be used when the desired results can be more efficiently attained by its use, and that a formula is nothing more than a statement in mathematical language of a method of computation already thought out and understood. The difficulties that may arise in this subject are not primarily mathematical. They are essentially a part of the necessarily difficult task of analyzing a statistical distribution.

The preparation of a book on mathematical statistics to appeal to scientific workers in fields ordinarily considered to be non-mathematical is essentially a matter of experimentation. It is the hope of the author that this book may stimulate interest in the methods of presenting statistical theory and in the more inclusive problem of making mathematical theory more widely available. Any suggestions or criticism of this presentation will be appreciated.

The Bibliography of Appendix II is inserted as a guide to advanced reading in the subject of mathematical statistics; the contributions of Prof. Pearson are to be noted especially.

It seems hardly necessary to refer to the debt which anyone who works in statistical theory must owe to Professor Karl Pearson. Because of his "Tables for Statisticians and Biometricians" the formulas of Appendix I are not given in more detail.

Professor James McMahon has given most generously of his time and interest. Whatever assistance this book may afford to the practical worker in statistics is in a large measure due to the influence of Professor Walter F. Willcox, whose critical insight into the limitations and the possibilities of statistical methods together with the originality and practical initiative which permeate his research and instructional work place all his students under obligations to him.

## CONTENTS.

## CHAPTER VII.



## CHAPTER VIII.



## CHAPTER IX.



## CHAPTER X.

## CHAPTER I.

### CURVE PLOTTING.

**Plotting the Data.** Let us plot the following data* of the monthly precipitation at Columbus for the year 1916:

| | | | | | |
|---|---|---|---|---|---|
| January, | 5.0 inches | May, | 4.8 inches | September, | 1.5 inches |
| February, | 1.5 inches | June, | 3.5 inches | October, | 1.8 inches |
| March, | 4.9 inches | July, | 0.7 inches | November, | 1.6 inches |
| April, | 2.3 inches | August, | 3.2 inches | December, | 3.6 inches |

A horizontal straight line is first drawn and at equal distances on this line twelve points are located, one for each month. On a vertical line erected at the point corresponding to the month of January equal intervals are laid off, one for each inch of precipitation; and these intervals are subdivided into tenths. The two series of points are called the **scales.** It is usual to designate the horizontal and the vertical scale lines by O — X and O — Y respectively, as in Figure 1.

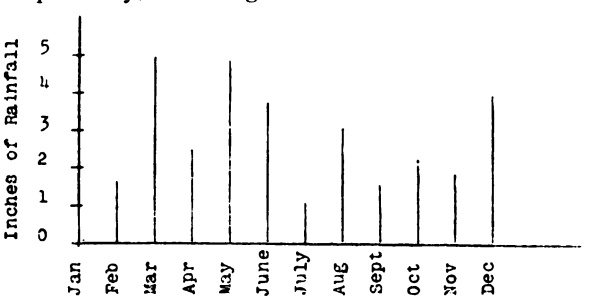


FIG. I. Monthly Precipitation at the Columbus Station for the year 1916.

The January precipitation is 5.0 inches. Place a dot above the January, or beginning point, at a height corresponding to 5.0 inches on the vertical scale. The next point is directly above the second or February point at a distance corresponding to 1.5 inches. Continuing in this way we locate a point for each month; the data is then said to be **plotted** or pictured point by point.

*Annual Meteorological Summary, U. S. Weather Bureau, Columbus, Ohio. 1917.

## Exercises.

1. Plot the following March precipitation data.*

| | | | |
|---|---|---|---|
| 1879 ..... 3.8 | 1889 ..... 0.7 | 1899 ..... 4.7 | 1909 ..... 2.7 |
| 1880 ..... 2.4 | 1890 ..... 5.6 | 1900 ..... 2.6 | 1910 ..... 0.3 |
| 1881 ..... 4.0 | 1891 ..... 4.6 | 1901 ..... 1.8 | 1911 ..... 2.4 |
| 1882 ..... 4.8 | 1892 ..... 2.2 | 1902 ..... 2.6 | 1912 ..... 4.6 |
| 1883 ..... 3.2 | 1893 ..... 1.9 | 1903 ..... 4.1 | 1913 ..... 8.1 |
| 1884 ..... 3.6 | 1894 ..... 1.8 | 1904 ..... 4.9 | 1914 ..... 2.5 |
| 1885 ..... 0.5 | 1895 ..... 1.2 | 1905 ..... 1.9 | 1915 ..... 1.2 |
| 1886 ..... 3.9 | 1896 ..... 3.0 | 1906 ..... 4.6 | 1916 ..... 4.9 |
| 1887 ..... 2.6 | 1897 ..... 5.5 | 1907 ..... 5.2 | |
| 1888 ..... 3.8 | 1898 ..... 7.0 | 1908 ..... 6.0 | |

2. Plot the following population data for the United States:

| | | |
|---|---|---|
| 1790 ..... 3,929,214 | 1840 ..... 17,069,453 | 1890 ..... 62,947,714 |
| 1800 ..... 5,308,483 | 1850 ..... 23,191,876 | 1900 ..... 75,994,575 |
| 1810 ..... 7,239,881 | 1860 ..... 31,443,321 | 1910 ..... 91,872,266 |
| 1820 ..... 9,638,453 | 1870 ..... 38,558,371 | |
| 1830 ..... 12,866,020 | 1880 ..... 50,155,783 | |

In plotting this data take the numbers to the nearest million.

**General Directions for the Laying off of Scales.** The object of any graphic representation of statistical data is to present a vivid picture and therefore a diagram too small or too large, or too wide or too narrow will not accomplish this purpose as efficiently as will a correctly proportioned diagram. This means that the widths of the horizontal and the vertical scale intervals must be carefully chosen in order to give the complete diagram the proper proportions.

In determining the widths of the intervals account must be taken of the nature of the statistical material. If the data is so inaccurate, for instance, that the measurements can be determined only to the nearest million it would be improper to divide the scale into intervals corresponding to thousands. The wealth of the country and the value of manufactured articles are examples of statistics which do not admit of close subdivision.

It is useless to have the scale intervals finer than the smallest difference which the eye can conveniently distinguish on the dia-

*Annual Meteorological Summary, U. S. Weather Bureau; Columbus, Ohio. 1917.

gram. This often means, even in the case of quite accurate material, that the figures of the data must be cut back; for instance in plotting population data for the United States one million may be the smallest numerical difference that can be pictured on an ordinary sized diagram.

Usually, as in Figure II, horizontal and vertical lines, called **coördinate lines,** are drawn to assist in carrying the divisions of the scales across the diagram. Care must be taken that these lines are lightly drawn and are not more numerous than is necessary.

**Connecting the Points.** The eye is assisted in passing across a diagram if the plotted points are connected by a curve. The curve may be either a series of broken straight lines joining the points or a continuous curve passing thru each point without sharp angles or abrupt changes in direction. Of the two methods the continuous curve is usually to be preferred because of the better appearance which it presents. In Figure II the points are connected by straight lines and in Figure III a continuous curve is drawn.

## Exercises.

3. Plot the curve of the 1916 rainfall at Columbus from the data of Exercise 1.

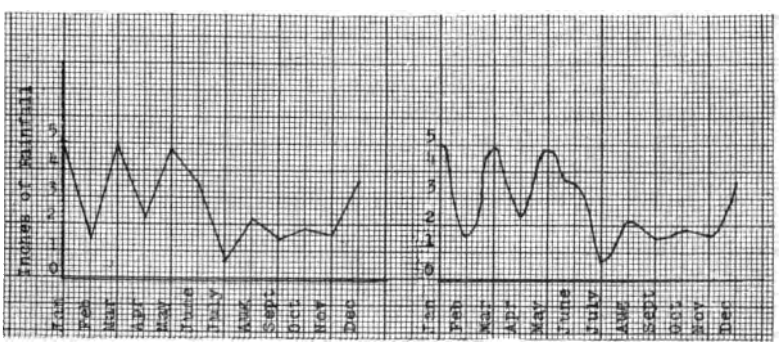

FIG. II. The Plotted Points of Monthly Temperatures connected by straight lines.

FIG. III. The Points of Fig. II connected by a continuous curve.

4. Plot the population curve from the data of Exercise 2.

## Directions for Plotting Curves.

1. The general arrangement of a diagram should be from left to right and from bottom to top.

2. Figures for the scales of a diagram should ordinarily be placed at the left and along the bottom.

3. Whenever practicable, the vertical scale should be so chosen that the zero line will appear on the diagram. When this is not done it is well to indicate that fact by a break in the diagram.

4. The zero lines must be sharply distinguished from the other coordinate lines of the diagram.

5. The curve must be carefully distinguished from the coordinate lines.

6. The data should accompany the diagram either in the form of a tabular statement or placed directly on the diagram. The latter method of presenting the original data can sometimes be effectively used, especially when the number of items is not large.

Underlying all rules for the construction of statistical diagrams is the general direction: *The diagram must be so arranged as to present the data most effectively.* Because of the great diversity of statistical material and of the wide variety of purposes for which data may be collected and presented it is not possible to lay down specific rules which are to be followed in every case. Whenever the vividness and accuracy of the statistical picture is not sacrificed by so doing, the conventional and generally accepted ways should be followed.

## Exercises.

5. Plot the following data of annual precipitation.*

| | | | |
|---|---|---|---|
| 1879 ..... 31.3 | 1889 ..... 28.5 | 1899 ..... 28.5 | 1909 ..... 36.6 |
| 1880 ..... 44.7 | 1890 ..... 50.7 | 1900 ..... 30 3 | 1910 ..... 34.8 |
| 1881 ..... 47 0 | 1891 ..... 42.1 | 1901 ..... 26.5 | 1911 ..... 43.4 |
| 1882 ..... 51.3 | 1892 ..... 33.5 | 1902 ..... 34.2 | 1912 ..... 29.6 |
| 1883 ..... 48.9 | 1893 ..... 38.1 | 1903 ..... 28.1 | 1913 ..... 40.9 |
| 1884 ..... 31.0 | 1894 ..... 29.5 | 1904 ..... 31.5 | 1914 ..... 31.2 |
| 1885 ..... 43.3 | 1895 ..... 30.7 | 1905 ..... 35.1 | 1915 ..... 39.9 |
| 1886 ..... 42.4 | 1896 ..... 40.5 | 1906 ..... 33.7 | 1916 ..... 34.4 |
| 1887 ..... 30.3 | 1897 ..... 41.2 | 1907 ..... 37.6 | |
| 1888 ..... 35.1 | 1898 ..... 41.3 | 1908 ..... 30.1 | |

Since the lowest number of inches is 26.5 it is better to make a break in the vertical scale, starting the working scale with, say, 25 inches.

* Report of Columbus Station, U. S. Weather Bureau.

6. Plot the following data of mean monthly temperatures.*

| | | | |
|---|---|---|---|
| 1879 ..... 53.1 | 1889 ..... 52.2 | 1899 ..... 53.2 | 1909 ..... 52.0 |
| 1880 ..... 53.6 | 1890 ..... 53.2 | 1900 ..... 53.8 | 1910 ..... 51.7 |
| 1881 ..... 54.2 | 1891 ..... 52.6 | 1901 ..... 51.8 | 1911 ..... 53.8 |
| 1882 ..... 53.4 | 1892 ..... 51.3 | 1902 ..... 52.1 | 1912 ..... 50.8 |
| 1883 ..... 51.8 | 1893 ..... 51.2 | 1903 ..... 52.0 | 1913 ..... 53.5 |
| 1884 ..... 52.5 | 1894 ..... 53.3 | 1904 ..... 50.2 | 1914 ..... 52.0 |
| 1885 ..... 49.1 | 1895 ..... 51.6 | 1905 ..... 51.5 | 1915 ..... 51.8 |
| 1886 ..... 50.3 | 1896 ..... 53.0 | 1906 ..... 52.7 | 1916 ..... 52.0 |
| 1887 ..... 52.5 | 1897 ..... 52.9 | 1907 ..... 50.8 | |
| 1888 ..... 51.0 | 1898 ..... 53.6 | 1908 ..... 53.5 | |

7. Plot the curve of Top Beef Cattle Prices from the following data:**

| | | | |
|---|---|---|---|
| 1891 ..... 7.15 | 1898 ..... 6.25 | 1905 ..... 7.00 | 1912 .....11.25 |
| 1892 ..... 7.00 | 1899 ..... 8.25 | 1906 ..... 7.60 | 1913 .....10.25 |
| 1893 ..... 6.75 | 1900 ..... 7.50 | 1907 ..... 8.00 | 1914 .....11.40 |
| 1894 ..... 6.40 | 1901 ..... 8.00 | 1908 ..... 8.40 | 1915 .....11.60 |
| 1895 ..... 6.60 | 1902 ..... 9.00 | 1909 ..... 9.50 | 1916 .....13.00 |
| 1896 ..... 6.50 | 1903 ..... 6.85 | 1910 ..... 8.85 | |
| 1897 ..... 6.00 | 1904 ..... 7.65 | 1911 ..... 9.35 | |

8. From the data of page 25 plot the 1916 beef cattle prices.
9. From the data of page 25 plot the 1895 beef cattle prices.

**The Title of a Diagram.** Each diagram must be provided with a brief and concise and yet accurate and comprehensive title. The title must cover all of the data and not merely a certain section of it and it must do this without being of undue length. A careful study of examples of titles is especially helpful in acquiring a notion of what constitutes a proper title.

All headings of columns must be clear and definite. The units of measurement of a scale must always be given; thus, "Precipitation *in inches,*" "Temperature *in degrees*".

Titles and headings have a better appearance when made in Roman characters than when made in script. In general the size of type in each heading or sub-title should correspond in size and prominence to its respective importance. Unless the lettering is skilfully done by hand it is better to use a typewriter even tho different sizes of letters cannot be secured by its use.

---

* Report of Columbus Station, U. S. Weather Bureau.

** Chicago Live Stock World, January 2, 1917.

## Exercises.

10. Study the titles and headings of the diagrams and tables of Vol. V, Report of the United States Census, 1910.

11. Study the titles shown in "Graphic Methods for Presenting Facts" by Willard C. Brinton.†

12. Study the titles and headings of the current issue of the Monthly Crop Reporter, Department of Agriculture.

In each of the following exercises construct a complete statistical diagram with the curve carefully drawn and an appropriate title designed for each.

13. The Land area of the United States exclusive of outlying possessions from Table 18, Vol. I, Report of the United States Census, 1910.

14. The population of Ohio from Table 10, same report.

15. Comparative Values of Inside Lots of Different Depths according to the Lindsay-Bernard system of valuation.

**The Lindsay-Bernard and Somers Valuation Schedule.***

| *Depth.* | *Lindsay-Bernard.* | *Somers.* | *Depth.* | *Lindsay-Bernard.* | *Somers.* |
|---|---|---|---|---|---|
| 5. | $9 | $14.35 | 85. | $82.0 | $93.33 |
| 10. | 15 | 25.00 | 90. | 84.2 | 95.60 |
| 15. | 21 | 32.22 | 95. | 86.2 | 97.85 |
| 20. | 27 | 41.00 | 100. | 88 | 100.00 |
| 25. | 33 | 47.90 | 105. | 89.6 | 102.08 |
| 30. | 38.5 | 54.00 | 110. | 91.1 | 104.00 |
| 35. | 44 | 59.20 | 115. | 92.5 | 105.78 |
| 40. | 49 | 64.00 | 120. | 93.8 | 107.50 |
| 45. | 54 | 68.45 | 125. | 95 | 109.50 |
| 50. | 58.5 | 72.50 | 130. | 96.1 | 110.50 |
| 55. | 63 | 76.20 | 135. | 97.2 | 111.80 |
| 60. | 67 | 79.50 | 140. | 98.2 | 113.00 |
| 65. | 70.6 | 82.61 | 145. | 99.2 | 114.50 |
| 70. | 73.9 | 85.60 | 150. | 100 | 115.00 |
| 75. | 76.9 | 88.30 | 175. | 103 | 119.14 |
| 80. | 79.6 | 90.90 | 200. | 105 | 122.00 |

16. Comparative Values of Inside Lots of Different Depths according to the Somers system.

17. The accumulated value of $1 at 10% compound interest:

| Year. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Amount .. | 1.00 | 1.10 | 1.21 | 1.33 | 1.46 | 1.61 | 1.77 | 1.95 | 2.14 | 2.36 |

† The Engineering Magazine, 1915, N. Y.

* The National Real Estate Journal, May, 1914.

18. The Average Yield per Acre for Wheat in the United States since 1866; Yearbook, Dep't of Agriculture.

19. Average Farm Price per bushel of Wheat in the United States since 1866; Yearbook, Dep't of Agriculture.

20. Per cent of Wheat Crop Exported since 1866; Yearbook, Dep't of Agriculture.

21. Total Production of Wheat to nearest 10 million bushels in the United States since 1866; Yearbook, Dep't of Agriculture.

22. Substitute the word Corn for Wheat in Exercises 17 to 20 and construct the curves.

23. Bank Clearings of the United States, excluding N. Y.

Bank Clearings of U. S. excluding N. Y. (in millions).

| | | | | | |
|---|---|---|---|---|---|
| 1883 | $14,209 | 1894 | $21,298 | 1905 | $50,087 |
| 1884 | 12,919 | 1895 | 23,507 | 1906 | 55,327 |
| 1885 | 13,170 | 1896 | 22,304 | 1907 | 57,994 |
| 1886 | 15,513 | 1897 | 23,895 | 1908 | 53,133 |
| 1887 | 17,566 | 1898 | 26,959 | 1909 | 62,249 |
| 1888 | 18,397 | 1899 | 33,416 | 1910 | 66,821 |
| 1889 | 20,280 | 1900 | 33,771 | 1911 | 67,857 |
| 1890 | 23,370 | 1901 | 39,152 | 1912 | 73,209 |
| 1891 | 23,198 | 1902 | 41,695 | 1913 | 75,181 |
| 1892 | 25,660 | 1903 | 43,239 | 1914 | 72,225 |
| 1893 | 23,049 | 1904 | 43,972 | | |

24. Percapita Imports of U. S.:

| | | | | | |
|---|---|---|---|---|---|
| 1860 | $11.25 | 1879 | $10.52 | 1897 | $10.32 |
| 1861 | 9.02 | 1880 | 13.88 | 1898 | 8.66 |
| 1862 | 5.79 | 1881 | 13.06 | 1899 | 10.68 |
| 1863 | 7.29 | 1882 | 14.36 | 1900 | 10.86 |
| 1864 | 9.30 | 1883 | 12.81 | 1901 | 11.34 |
| 1865 | 6.87 | 1884 | 11.48 | 1902 | 12.30 |
| 1866 | 12.26 | 1885 | 10.49 | 1903 | 12.42 |
| 1867 | 10.23 | 1886 | 11.57 | 1904 | 12.71 |
| 1868 | 9.94 | 1887 | 12.09 | 1905 | 14.24 |
| 1869 | 11.60 | 1888 | 12.11 | 1906 | 15.69 |
| 1870 | 11.97 | 1889 | 12.58 | 1907 | 16.29 |
| 1871 | 14.47 | 1890 | 13.15 | 1908 | 12.54 |
| 1872 | 16.15 | 1891 | 12.96 | 1909 | 16.28 |
| 1873 | 14.27 | 1892 | 12.91 | 1910 | 10.94 |
| 1874 | 13.13 | 1893 | 11.68 | 1911 | 16.32 |
| 1875 | 11.43 | 1894 | 9.97 | 1912 | 19.04 |
| 1876 | 9.47 | 1895 | 11.60 | 1913 | 18.47 |
| 1877 | 10.37 | 1896 | 9.66 | 1914 | 18.14 |
| 1878 | 9.07 | | | | |

Note that the data of the two preceding exercises shows a decided periodicity or wave-like nature.

**More than One Curve on the Same Diagram.** For the purpose of comparing different curves it is often convenient to plot two or more curves on the same diagram. For instance, simultaneous variations in the prices of wheat and corn can be observed to good advantage, when the two curves are brot together on the same diagram and constructed to the same scales. The chief disadvantage of this method of comparing curves lies in the resulting complexity of the diagram. If the diagrams are constructed on thin paper and the lettering and curves are made heavy the different curves when made on separate sheets can be readily compared by adjusting one sheet of paper above the other.

### Exercises.

25. Compare the rainfall curve of Exercise 5 with the temperature curve of Exercise 6. To what extent do the two curves vary in the same directions? What conclusions can be drawn as to the tendency for the amount of rainfall to depend on the temperature?

26. Compare the two systems of real estate valuation of Exercises 15 and 16.

27. Give a comparative interpretation of the curves of Exercises 18 and 19. Why should they not be expected to follow exactly the same general course?

28. Discuss as in Exercise 27 curves of prices and yield per acre of corn.

29. Compare the curves of Exercises 21 and 23.

**Coördinates.** It is convenient to have a standardized notation for the horizontal and vertical scales. The horizontal line is denoted by O-X and called the **axis of abscissas** or simply **the X-axis.** The vertical line is denoted by O-Y and called the **axis of ordinates** or the **Y-axis.** The point where the two lines meet is the **origin of coördinates.** Distances along the X-axis are spoken of as **x distances or x coördinates,** and those along the Y-axis as **y distances,** or **y coördinates.** Thus in the precipitation data of page 7, the origin is at January, 1879, and the values of X differ by intervals of one month, while the unit interval for Y is one inch.

**Logarithmic Curves.** Whenever the data seems to exhibit a uniform *rate* of increase or whenever it is desired to study the *relative* changes rather than the *actual* changes in the

data, a logarithmic curve may be of service.* **A logarithmic curve** is obtained by taking the logarithms of the measurements and using these logarithms as vertical distances or ordinates. Since multiplying two numbers adds their logarithms,a constant ratio or rate will appear in the logarithmic diagram as a constant addition. Hence if there is a constant rate in the data the logarithmic curve will be a straight line. Whether the rate is constant or not, curves of this type are of value for comparing different rates. However, if the rate is not approximately constant considerable familiarity with logarithms is necessary if the comparative differences are to be correctly interpreted.

### Exercises.

30. Plot the logarithmic curve of the data of Exercise 17.
31. Plot the logarithmic curves of the data of Exercises 15 and 16.
32. Plot the logarithmic curve of the Chicago Top Beef Cattle Prices.

**Cumulative Curves.** All the preceding curves show the respective values for each interval of the horizontal axis, as the production of wheat for each year since 1866 is shown by the curve of Exercise 21. Now if it is desired to construct a curve exhibiting at each year the total production of wheat since 1866, the amount of each year's production is added to that of all the preceding years and the resultant cumulative sums plotted. In this way a curve is obtained which starts at the lower left hand corner and proceeds in a diagonal direction across the diagram. It is called a **cumulative** curve. The values to be plotted will be, in the case of the cumulative curve of wheat production, 150,000,000, 360,000,000, 580,000,000, 840,000,000, etc.

### Exercises.

33. Plot the cumulative curve of wheat production.
34. Plot the cumulative curve of corn production and compare with the curve of Exercise 33.
35. Of what significance is the slope of a cumulative curve?

---

*See "The Ratio Curve," Fisher. Quarterly Publications American Statistical Association, June, 1917.

## CHAPTER II.

### CURVE PLOTTING—(*Continued.*)

**Interpolation.** The curves of the preceding chapter were drawn for the purpose of connecting the plotted points in order to assist the eye in following the course of the data across the diagram. However, other uses can be made of a statistical curve.

At the beginning of Chapter I the data of monthly precipitation is given. What was the weekly precipitation? The Chicago Top Beef Cattle monthly prices are given under Exercise 7, Chapter I. What were the weekly prices during the period covered by that data? The population of the United States is given for ten-year intervals. What has been the population from year to year? These are essentially questions of **interpolation**, that is, of estimating values lying between the given values.

The method of obtaining intermediate values from the curve consists merely of measuring on the vertical scale the height of the curve at the required point. Thus with the population curve of Exercise 4, Chapter I, which is constructed from the decennial census reports, the population for the year 1906 is given by the height of the curve above the 1906 point on the horizontal scale.

#### Exercises.

1. Estimate the Top Beef Cattle Prices for each week in February 1916, from the monthly data of Exercise 7, Chapter 1.

2. Estimate the values of inside lots for the fraction of a foot, say 67.5 feet, from the data of Exercise 16 of the preceding chapter.

3. What is, according to the data of Exercise 17 of Chapter I, the compound amount of $1 for 7.5 years at 10%?

This method of interpolating makes an estimated value depend on the two consecutive given values which inclose it. But the increase in population during a decade may have occurred almost entirely during the last years of the period and

yet the shape of the curve when drawn merely to connect the ten-year points may give no hint of this irregularity of increase. The temperature for one month may have no connection with that of the preceding month and hence the curve between the points, depending as it does on the two non-related values can hardly be expected to give the actual temperature for an intermediate week or day. If the price of wheat for the year 1905 is omitted can it be reliably estimated by drawing the curve from the years 1904 and 1906 and then interpolating for the missing year?

It must be apparent therefore that a curve which passes thru a series of more or less non-related points can be of little value in interpolation and that the problem of interpolation is essentially one of determining by some means or other the general course of the data and then estimating the intermediate values in conformity with this general trend. The values obtained in this way are the most probable values; accidental variations which bear no relation to the underlying tendencies can not be so estimated; in fact such variations can not be estimated or predicted by any means.

**The Smoothing of a Curve.** The curves of Chapter I, drawn as they are thru each point, preserve all the variations whether they are fundamentally essential or due merely to the presence of accidental influences. The curve of mean monthly temperatures, Exercise 6 of the preceding chapter, shows distinct seasonal variations in temperature—higher temperatures in summer and lower in winter. Along with these essentially significant changes are fluctuations apparently accidental as, in one year June is warm and in another relatively cool; sometimes January is warmer than February and sometimes the reverse is true.

To represent a general movement or trend the curve must be drawn without abrupt changes in direction and must sweep *among* the points rather than necessarily thru each point. Since such a **smoothed curve**, as it is called, depends on the general or collective characteristics of the data the drawing of it must be based on collective properties of the measurements. One pertinent general property has just been stated; namely, that the curve must be smooth, that is, not have abrupt

changes in direction. This property expresses the statistical assumption that the significant variations are fairly uniform from value to value and not capricious or arbitrary. A second assumption, which is presently discussed, is that certain areas are relatively stable and unchanging.

**Smoothing by Inspection.** The smoothing of a curve may be based on a study of the data and made a matter of the skill and experience of the statistician without the assistance of definitely stated assumptions or properties. The curve is then said to be smoothed by **inspection.**

In smoothing a curve the first step is to study the data carefully. Without such an investigation into the probable sources and extent of the irregularities and fluctuations one cannot hope to know what irregularities to smooth out and what to leave in. A curve cannot be reliably smoothed by a statistician who does not know the data thoroly. On the basis of the information gained by this study a preliminary curve should then be drawn freehand among the points. By successive erasures and redrawings the finished curve can gradually be arrived at. Thus a curve showing the long time movements in the price of wheat will pass above some points and below others and how much the curve should miss any point can not be determined without a knowledge of financial conditions, yields, etc.

The inspection method of smoothing a curve is often sufficiently accurate for all practical purposes, especially when done by a statistician of experience and especially when there is a considerable element of inaccuracy inherent in the data. Its disadvantage lies obviously in the fact that no two smoothings of the same curve will be exactly alike; the method is essentially tentative and personal.

In any event a rough preliminary draft of the curve should be made by inspection before proceeding to apply more refined methods.

### Exercises.

4. Smooth the illustrative data at the beginning of Chapter I.
5. Smooth the data of the population of the United States as given in Exercise II, Chapter I.
6. Smooth the data of annual rainfall of Exercise 5, Chapter I.
7. Smooth the data of Exercises 18, 19, 20 and 21 of Chapter I.

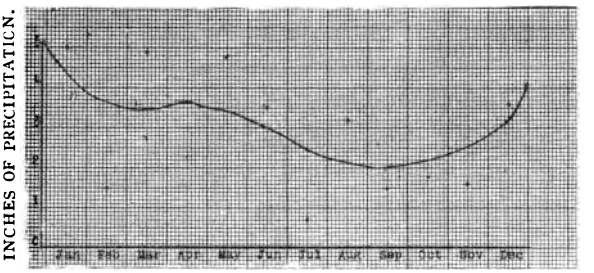


FIG. IV. The Smooth Curve of Monthly Precipitation at Columbus, 1916.

**The Preservation of Areas.** In the illustrative data at the beginning of Chapter I the precipitation of 4.9 inches in March is the total precipitation for the whole month. With a base of one unit, then a rectangle of height 4.9 will have an area equal to the total precipitation. Likewise the rectangle on the July unit as a base will have an area equal to 0.7, which is the July precipitation. The prices of Exercise 7, Chapter I, can in a similar manner be represented by rectangles with heights equal to the respective prices and with unit bases. The population data of Exercise 2 of the same chapter may be represented by rectangles which are not adjacent and have nine rectangles omitted between successive census years.

After the curve is smoothed each rectangle will be altered so as to have a curved top. The total area under the finished curve will then be the sum of the areas of the modified rectangles. **The First Rule of Preservation of Areas** is that the *curve should be so smoothed that the total area under the resulting curve is equal to the sum of the areas of the original rectangles.* Since, for instance, the monthly precipitation is made up of the sum of the daily precipitations it is likewise reasonable to assume that the monthly sum is more stable than is the daily or weekly and hence we have the **Second Rule of the Preservation of Areas**; namely, *where possible, the areas of the individual rectangles are to remain unchanged.* This can be done by adding to and subtracting from each rectangle an equal sum.

Within the requirement that the curve must be free from abrupt changes in direction the two preceding working rules furnish a fairly comprehensive basis for the smoothing of

statistical data. In later chapters more detailed rules will be discussed and applied. However, for most data the present rules are sufficient.

As explained for the precipitation data a definite statistical meaning can usually be found for the rectangles. Even when a significance is with difficulty ascribed to the rectangles they should be drawn and the same rules applied to the smoothing as before. The method is in such cases justified wholly by its practical convenience.

In the illustrative plotting, at the beginning of Chapter I, of the data of monthly precipitation at Columbus for the year 1916, the vertical scale was laid off on a line thru the January point. In constructing the rectangles for smoothing, it is convenient to have the January and other perpendiculars at the middle of the respective intervals in order that there may be a half unit's space at the left of the beginning point. The zero point on the horizontal scale is then at the beginning of the first interval and the vertical distance for the first point is taken not on the vertical scale line but perpendicularly above the mid-point of the interval. Whenever the curve is to be smoothed the scale is marked off in this way; ordinarily the method of Chapter I is employed where the curve is not to be smoothed.

The following diagram illustrates the application of the rectangle method of smoothing to the monthly precipitation data.

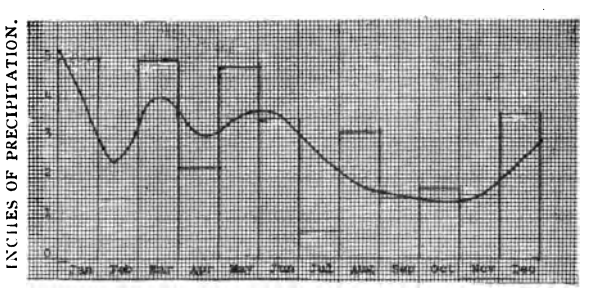


FIG. V. The Rectangle Method of Smoothing the Monthly Precipitation data for Columbus in 1916.

## Exercises.

8. Construct the smoothed curve of prices from the data of Exercise 7, Chapter I.

9. Do the same for the data of Exercise 5 and of Exercise 6, of the same chapter.

10. Do the same for the data of Exercises 18, 19, 20 and 21 of the same chapter.

11. Do the same for the data of Exercise 22 of the same chapter.

12. Can the rules of *permanence of areas* be applied effectively to the drawing of the curve for the data of Exercise 17 of the preceding chapter? Why? To the data of Exercises 15 and 16 of the same chapter?

13. In drawing the smooth curve of decennial census population it is advisable to alter the original data very slightly, if at all. Discuss.

A common way of drawing this curve is to connect the ten year points by a series of straight lines and then round out the angles where the lines intersect. This assumes a uniform annual increase during the decade — an assumption which may or may not be true.

14. The statistical significance of the rectangles has been discussed for the precipitation data. Develop the corresponding explanation for the decennial census data.

15. Show that in the data of Exercises 15, 16 and 17 of the preceding chapter the rectangles are not significant.

**The Adjusted Data; Interpolation.** Since in general it is impossible to preserve exactly the area of each rectangle the process of smoothing will lead to values differing from those of the original data. Consequently, the data is said to be **adjusted** or **graduated** or **smoothed** by means of the curve. In accordance with the reasoning at the beginning of this chapter the adjusted values are to be taken as giving a more significant idea of the true trend of the data than does the original data. It is evident that we have here the solution to the problem of interpolation. Therefore, the **rule for interpolation is:** *to obtain the value at any point on the horizontal scale measure the corresponding ordinate of the smoothed curve, or measure the proper area under that curve.* Thus the rainfall during the first week in June is obtained by measuring the area under the curve on the first one-fourth of the June base unit.

**Test of a Graduation.** The extent to which smoothing preserves the areas of the individual rectangles is often taken as a test of the appropriateness of the smoothing or graduation. The smoothed curve is said to **fit** the data and the term **"goodness of fit"** is used to denote the appropriateness of the methods used in the process of constructing the smooth curve. The goodness of fit is then measured by the extent to which the areas of the individual rectangles are preserved. In

applying this test two columns of numbers are set down, in one the original values and in the other the adjusted values. The differences are then taken and studied. Other conditions being equal the smoothing with the smallest differences is the best, tho the judging of goodness of fit is largely a matter of experience.

### Exercises.

16. Discuss the goodness of fit of each of the curves smoothed in the preceding exercises.

17. What is the best estimate on the basis of the data of page 25 of the Top Beef Cattle Prices for the first week in February, 1916?

Note that in this data the rectangles have no special statistical significance.

18. From the data of Exercise 23 of the preceding chapter what is the best estimate of the bank clearings in the United States for the first half of the year 1908?

19. What is the significance of the rectangles in the case of the data of Exercises 14, 15, 16 of Chapter 1?

20. In drawing the curves of Exercise 19 should the values be adjusted? Are these curves drawn by a process of smoothing?

**Determining the General Trend of the Data.** The characteristics of a movement over a number of years can be determined from the smoothed curve. Thus the general upward trend of prices during the years 1897 to 1917 is shown by the rise of the curve.

Perhaps the best way to picture a general movement in the data is to draw a straight line, or more than one straight line where there seems to be more than one distinct movement, to fit the data. That is, to smooth the data with a straight line. With data not conforming closely to a straight line there is likely to be some uncertainty in the exact location of the straight line or lines but since the lines are but the pictures of the ideas of general increases or decreases the uncertainty is neither greater nor less than is the uncertainty in the ideas of the general movements themselves. The difficulty, in reality, is due to a lack of information regarding the data. The methods of Chapter X are of much service in this connection.

### Exercises.

21. During the last 37 years has there been an appreciable increas or decrease in the precipitation at the Columbus Station?

22. During the same time has there been a decided upward c downward movement in temperatures at the same place?

**Periodic Data.** In smoothing and determining the general trend of data care must be taken that the data is not smoothed to conform to a straight line when there is an inherent periodicity in the material. The data of Exercises 23 and 24 of Chapter I exhibit significant tendencies for the values to be high for a few years and then consistently lower for a few years and then higher, and so on, thru more or less regular and uniform cycles. In smoothing such data the ideal should be to determine a uniform cycle and then smooth the data into the curve made up of the determined cycles. The problem of smoothing such data is complicated by the fact that the curve in addition to being composed of a series of similar loops or arches also has a tendency to rise or fall. Thus the imports of the U. S. have increased on the whole during the last 50 years tho there have been increases and decreases following each other in fairly regular periods.

### Exercises.

23. Smooth the data of Bank Clearings as given in Exercise 23 of the preceding chapter.

24. Smooth the data of Imports as given in Exercise 24 of the preceding chapter.

25. To what extent has there been a tendency for bank clearings and for imports to increase during the period covered by the given data?

26. Discuss the periods in the yield per acre of wheat in the U. S.

27. Do the same for the production of wheat.

28. Summarize the uses and advantages of the smooth curve as compared with the curve which passes exactly thru each point.

## CHAPTER III.

### FREQUENCY CURVES.

**Definitions.** The following data of the measures of heights of 750 students* may be taken for purpose of illustration.

The measurements are classified to show the number of individuals for each inch of height.

| *Height.* | *Number.* | *Height.* | *Number.* |
|---|---|---|---|
| 61 | 2 | 68 | 126 |
| 62 | 10 | 69 | 109 |
| 63 | 11 | 70 | 87 |
| 64 | 38 | 71 | 75 |
| 65 | 57 | 72 | 23 |
| 66 | 93 | 73 | 9 |
| 67 | 106 | 74 | 4 |
| | | | 750 |

TABLE I.

Height, the **attribute** or **characteristic** here under consideration, is in this table measured to the nearest inch, giving a **group** or **class interval** of one inch. A class interval or class is ordinarily **designated** by the value of its middle measurement, and the **class limits** are located on either side at a half unit's distance from this mid-value. All individuals, for instance, with height between 67.5 and 68.5 belong to class 68; here the limits are 67.5 and 68.5 and the class is designated by the number 68. Instead of using 61, 62, 63, etc., as class numbers. the classes may be simply numbered 1, 2, 3, etc., and these numbers used as class numbers. Again, the classes may be numbered in both ways from some point within the range, as 68. This would give class numbers as follows: $-7$, $-6$, $-5$, $-4$, $-3$, $-2$, $-1$, 0, $+1$, $+2$, etc.

The objects measured or enumerated are referred to as **variates** or simply as **individuals**.

* Records of physical measurements at Ohio State University Gymnasium, Freshman class, 1913.

The size or **frequency** of a class is the number of individuals within that class, and the **total frequency** is the sum of all the class frequencies. The table as a whole constitutes a **frequency distribution** of height, and shows the number of times each class occurs.

To illustrate the method of constructing a frequency distribution let us take the following data: *

**Chicago Monthly Top Beef Cattle Prices.**

| Year. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1916..... | $9.85 | $9.75 | $10.05 | $10.00 | $10.90 | $11.50 | $11.30 | $11.50 | $11.50 | $11.60 | $12.40 | $13.00 |
| 1915..... | 9.70 | 9.50 | 9.15 | 8.90 | 9.65 | 9.95 | 10.40 | 10.50 | 10.50 | 10.60 | 10.55 | 11.60 |
| 1914..... | 9.50 | 9.75 | 9.75 | 9.55 | 9.60 | 9.45 | 10.00 | 10.90 | 11.05 | 11.00 | 11.00 | 11.40 |
| 1913..... | 9.50 | 9.25 | 9.30 | 9.25 | 9.10 | 9.20 | 9.20 | 9.25 | 9.50 | 9.75 | 9.85 | 10.25 |
| 1912..... | 8.75 | 9.00 | 8.85 | 9.00 | 9.40 | 9.60 | 9.85 | 10.65 | 11.00 | 11.05 | 11.00 | 11.25 |
| 1911..... | 7.10 | 7.05 | 7.35 | 7.10 | 6.50 | 6.75 | 7.35 | 8.20 | 8.25 | 9.00 | 9.25 | 9.35 |
| 1910..... | 8.40 | 8.10 | 8.85 | 8.65 | 8.75 | 8.85 | 8.60 | 8.50 | 8.50 | 8.00 | 7.75 | 7.55 |
| 1909..... | 7.50 | 7.15 | 7.40 | 7.15 | 7.30 | 7.50 | 7.65 | 8.00 | 8.50 | 9.10 | 9.25 | 9.50 |
| 1908..... | 6.40 | 6.25 | 7.50 | 7.40 | 7.40 | 8.40 | 8.25 | 7.90 | 7.85 | 7.65 | 8.00 | 8.00 |
| 1907..... | 7.30 | 7.25 | 6.90 | 6.75 | 6.50 | 7.10 | 7.50 | 7.60 | 7.35 | 7.45 | 7.25 | 6.35 |
| 1906..... | 6.50 | 6.40 | 6.35 | 6.35 | 6.20 | 6.10 | 6.50 | 6.85 | 6.95 | 7.30 | 7.40 | 7.90 |
| 1905..... | 6.35 | 6.45 | 6.35 | 7.00 | 6.85 | 6.35 | 6.25 | 6.50 | 6.50 | 6.40 | 6.75 | 7.00 |
| 1904..... | 5.90 | 6.00 | 5.80 | 5.80 | 5.90 | 6.70 | 6.65 | 6.40 | 6.55 | 7.00 | 7.30 | 7.65 |
| 1903..... | 6.85 | 6.15 | 5.75 | 5.80 | 5.65 | 5.15 | 5.65 | 6.10 | 6.15 | 6.00 | 5.85 | 6.00 |
| 1902..... | 7.75 | 7.35 | 7.40 | 7.50 | 7.70 | 8.50 | 8.85 | 9.00 | 8.85 | 8.75 | 7.40 | 7.75 |
| 1901..... | 6.15 | 6.00 | 6.25 | 6.00 | 6.10 | 6.55 | 6.40 | 6.40 | 6.60 | 6.90 | 7.25 | 8.00 |
| 1900..... | 6.60 | 6.10 | 6.05 | 6.00 | 5.85 | 5.90 | 5.85 | 6.20 | 6.15 | 6.00 | 6.00 | 7.50 |
| 1899..... | 6.30 | 6.25 | 5.90 | 5.85 | 5.75 | 5.75 | 6.00 | 6.65 | 6.90 | 7.00 | 7.15 | 8.25 |
| 1898..... | 5.50 | 5.85 | 5.80 | 5.50 | 5.50 | 5.35 | 5.65 | 5.75 | 5.85 | 5.90 | 6.25 | 6.25 |
| 1897..... | 5.50 | 5.40 | 5.65 | 5.50 | 5.45 | 5.30 | 5.25 | 5.50 | 6.00 | 5.40 | 6.00 | 5.65 |
| 1896..... | 5.00 | 4.75 | 4.75 | 4.75 | 4.55 | 4.65 | 4.60 | 5.00 | 5.30 | 5.30 | 5.45 | 6.50 |
| 1895..... | 5.80 | 5.80 | 6.60 | 6.40 | 5.25 | 6.00 | 6.00 | 6.00 | 6.00 | 5.60 | 5.00 | 5.50 |

The width of the classes must be first determined. It would be possible to have a class for each quotation but it would be found highly inconvenient. The error introduced by the grouping of the measurements, the quotations in this case, is ordinarily of no practical significance. A general rule in determining the width of the classes, and hence of the number of classes, is to make as wide classes as is practically feasible—the number of classes is perhaps most often from ten to twenty. In this case the width is taken as fifty cents and the limiting quotations of each class are included in the class.

The data is examined and a score made for each occurrence of the class. Thus Class I with the range 450-499 appears Feb., Mar., Apr., May, June, July, 1896; as an occurrence is observed

* Yearbook, Chicago Live Stock World, 1917.

a mark or score is made—six in all. After the scoring is completed the frequency of each class, that is, the number of tallies or scores for each class, is noted and written in a column.

The frequency distribution just obtained *shows the number of times each price-class has occurred during the last twenty-one years.*

## Exercises.

1. Construct a frequency table from the Top Beef Cattle Prices using class intervals of twenty-five cents and compare with the distribution obtained when the class interval is fifty cents.

2. From the following table of Mean Monthly Temperatures at the Columbus Station construct a frequency table with a class width of five degrees.

| Year. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1878..... | .... | .... | .... | .... | .... | .... | 78.6 | 74.0 | 65.9 | 53.5 | 43.1 | 26.2 |
| 1879..... | 25.6 | 28.9 | 41.6 | 50.3 | 64.1 | 71.4 | 78.4 | 71.2 | 61.4 | **62.4** | 43.8 | 37.2 |
| 1880..... | **43.8** | 38.8 | 40.8 | 53.8 | 68.8 | 72.8 | 75.1 | 74.2 | 65.4 | 52.2 | **32.8** | **24.6** |
| 1881..... | 24.2 | 29.2 | 36.8 | 47.6 | 67.8 | 69.7 | 78.6 | 75.8 | **74.6** | 60.5 | 44.4 | 40.6 |
| 1882..... | 32.6 | **41.8** | 44.8 | 51.0 | 57.4 | 69.9 | 71.8 | 72.2 | 66.2 | 59.0 | 42.6 | 32.0 |
| 1883..... | 27.1 | 34.3 | 35.3 | 51.0 | 60.3 | 70.7 | 74.1 | 70.2 | 64.0 | 55.6 | 44.7 | 34.8 |
| 1884..... | 20.0 | 37.2 | 39 2 | 49.4 | 61.7 | 72.9 | 73.8 | 72.8 | 71.0 | 58.9 | 41.4 | 32.2 |
| 1885..... | 22.9 | **19.4** | **30.1** | 50.0 | 61.2 | 69.0 | 76.6 | 71.1 | 64.0 | 61.4 | 40.9 | 32.4 |
| 1886..... | 23.4 | 26.8 | 39.0 | 54.7 | 62.8 | 67.3 | 72.2 | 71.6 | 65.8 | 54.4 | 38.8 | 27.2 |
| 1887..... | 26.8 | 36.4 | 37.3 | 50.8 | 67.6 | 71.2 | 79.6 | 70.8 | 64.0 | 51.4 | 41.7 | 32.8 |
| 1888..... | 26.8 | 33.0 | 36.5 | 51.2 | 60.6 | 71.6 | 73.2 | 71.4 | **61.3** | 48.7 | 44.1 | 34.2 |
| 1889..... | 34.2 | 26.4 | 42.2 | 61.8 | 61.4 | 67.7 | 74.1 | 70.2 | 63.8 | 49.0 | 41.2 | **44.6** |
| 1890..... | 39.1 | 40.6 | 35.2 | 52.3 | 60.0 | 74.6 | 73.6 | 70.2 | 63.1 | 53.8 | 44.6 | 31.8 |
| 1891..... | 33.0 | 36.8 | 34.8 | 52.9 | 57.6 | 72.4 | **70.0** | 71.0 | 69.4 | 52.8 | 40.3 | 40.0 |
| 1892..... | 24.0 | 35.7 | 36.0 | 49.4 | 61.0 | 74.2 | 74.0 | 73.0 | 66.4 | 53.6 | 38.2 | 30.0 |
| 1893..... | **18.8** | 30.8 | 39.5 | 51.6 | 59.4 | 71.4 | 76.4 | 71.8 | 66.7 | 55.1 | 40.0 | 33.0 |
| 1894..... | 34.7 | 29.4 | 46.2 | 51.5 | 60.6 | 72.4 | 75.2 | 72.2 | 69.1 | 54.8 | 39.0 | 34.6 |
| 1895..... | 24.1 | 21.0 | 36.7 | 53.2 | 62.8 | **74.9** | 73.8 | 75.6 | 71.2 | **48.2** | 42.4 | 34.9 |
| 1896..... | 30.8 | 31.8 | 33.7 | **58.6** | **69.7** | 70.8 | 74.4 | 72.9 | 63.2 | 50.4 | 45.2 | 36.4 |
| 1897..... | 26.4 | 34.0 | 43.1 | 50.6 | 57.9 | 69.8 | 77.2 | 71.1 | 68.8 | 59.8 | 42.7 | 33.7 |
| 1898..... | 33.2 | 31.5 | 46.3 | 48.6 | 62.7 | 73.8 | 77.7 | 75.0 | 69 8 | 55.0 | 39.6 | 30.0 |
| 1899..... | 29.4 | 22.8 | 38.3 | 55.2 | 65.4 | 73.4 | 76.2 | 75.8 | 65.7 | 59.2 | 45.3 | 31.1 |
| 1900..... | 32.8 | 26.8 | 34.5 | 51.8 | 65.4 | 71.9 | 76.2 | **78.5** | 71.2 | 62.2 | 42.4 | 32.5 |
| 1901..... | 30.4 | 23.5 | 40.6 | 48.5 | **61.1** | 73.4 | **79.9** | 74.6 | 66.6 | 55.7 | 38.6 | 28.7 |
| 1902..... | 28.8 | 23.2 | 42.7 | 50.1 | 65.0 | 68.2 | 75.2 | 70.6 | 65.2 | 56.2 | **49.8** | 30.7 |
| 1903..... | 28.3 | 31.8 | 47.8 | 51.2 | 65.8 | **66.0** | 74.5 | 73.0 | 67.6 | 55.4 | 38.4 | **24.6** |
| 1904..... | 22.8 | 24.8 | 40.9 | 45.0 | 62.1 | 69.6 | 73.4 | 71.0 | 67.1 | 54.0 | 42.4 | 29.0 |
| 1905..... | 24.0 | 22.0 | 44.4 | 50.4 | 62.3 | 70.8 | 74.9 | 73.3 | 67.1 | 53.7 | 40.6 | 34.2 |
| 1906..... | 36.6 | 28.4 | 32.0 | 54.7 | 62.8 | 71.0 | 73.2 | 75.7 | 70.0 | 53.0 | 42.4 | 32.7 |
| 1907..... | 33.5 | 27.2 | 46.8 | **43.0** | **55.7** | 67.2 | 73.9 | 71.0 | 66.7 | 50.0 | 40.2 | 34.6 |
| 1908..... | 30.0 | 29.6 | 44.5 | 51.7 | 64.0 | 70.6 | 75.6 | 72.7 | 70.4 | 55.6 | 42.8 | 34.3 |
| 1909..... | 32.8 | 36.2 | 38.1 | 50.1 | 59.8 | 71.8 | 72.0 | 73.4 | 64.1 | 49.9 | 49.6 | 25.8 |
| 1910..... | **18.2** | **26.2** | **50.1** | 52.6 | 57.1 | 68.4 | 75.2 | **73.4** | 67.6 | 57.8 | 37.0 | 26.5 |

| Year. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1911..... | 33.5 | 35.4 | 38.3 | 49.0 | 68.6 | 72.8 | 75.7 | 73.9 | 68.8 | 54.1 | 38.2 | 37.4 |
| 1912..... | 19.2 | 23.4 | 34.2 | 43.4 | 64.2 | 68.2 | 74.9 | 70.7 | 68.0 | 56.2 | 42.6 | 34.4 |
| 1913..... | 36.8 | 27.2 | 40.8 | 50.4 | 61.7 | 71.8 | 76.5 | 75.4 | 65.4 | 54.4 | 45.8 | 35.5 |
| 1914..... | 34.0 | 23.6 | 36.5 | 50.7 | 63.8 | 72.6 | 75.9 | 74.0 | 64.8 | 57.7 | 42.9 | 27.6 |
| 1915..... | 27.8 | 36.0 | 34.0 | 56.3 | 58.2 | 68.0 | 73.0 | **68.2** | 68.0 | 56.7 | 44.4 | 31.0 |
| 1916..... | 36.0 | 26.8 | 35.5 | 49.9 | 62.6 | 65.9 | 78.6 | 75.8 | 63.9 | 55.2 | 43.4 | 30.4 |

3. Construct the frequency table of the following data of monthly precipitation at the Columbus Station. Take one-half inch for the class width.

| Year. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1878..... | .... | .... | .... | .... | .... | .... | 3.58 | 5.00 | 2.84 | 3.17 | 3.06 | 3.88 |
| 1879..... | 1.66 | 1.43 | 3.77 | 0.92 | 2.09 | 2.68 | 3.67 | 4.64 | 2.33 | **0.28** | 3.52 | 4.29 |
| 1880..... | 4.49 | 1.70 | 2.42 | 5.08 | 3.21 | 3.30 | 4.86 | 6.95 | 1.80 | 2.35 | 4.54 | 3.98 |
| 1881..... | 2.25 | 4.44 | 4.01 | 2.04 | 2.00 | 4.02 | 5.33 | 2.09 | 1.54 | **8.64** | 5.35 | **5.23** |
| 1882..... | 4.69 | 5.94 | 4.76 | 4.87 | **9.59** | 6.01 | 2.62 | 3.14 | 2.91 | 2.44 | 2.05 | 2.23 |
| 1883..... | 3.20 | 6.18 | 3.20 | 2.85 | 6.38 | 4.25 | 3.75 | 2.54 | 2.43 | 6.11 | 3.87 | 4.12 |
| 1884..... | 2.25 | 4.95 | 3.59 | 2.11 | 3.79 | 2.59 | 2.16 | 0.70 | 3.46 | 1.66 | 0.99 | 2.77 |
| 1885..... | 3.75 | 2.39 | 0.53 | 4.61 | 5.83 | 5.08 | 3.28 | 5.90 | 2.84 | 3.11 | 3.08 | 1.85 |
| 1886..... | 4.36 | 1.26 | 3.90 | 3.57 | 7.67 | 2.69 | 4.17 | 2.44 | 3.61 | 1.13 | 4.18 | 3.41 |
| 1887..... | 2.35 | 6.48 | 2.56 | 3.44 | 2.97 | 2.82 | 1.45 | 2.21 | 1.35 | 0.30 | 2.45 | 1.87 |
| 1888..... | 3.73 | 1.30 | 3.79 | 1.53 | 3.89 | 1.62 | 5.81 | 4.34 | 0.91 | 3.77 | 3 26 | 1.11 |
| 1889..... | 3.37 | 1.06 | 0.66 | **0.83** | 3.92 | 2.77 | 2.94 | 1.59 | 3.34 | 1.83 | 3.83 | 2.36 |
| 1890..... | 5.73 | 6.12 | 5.63 | 4.32 | 5.12 | 4.95 | 1.80 | 2.75 | **7.13** | 3.02 | 1.97 | 2.19 |
| 1891..... | 2.84 | 5.42 | 4.64 | 2.26 | 2.73 | 4.98 | 4.69 | 2.64 | 1.05 | 2.94 | 5.44 | 2.42 |
| 1892..... | 2.21 | 3.35 | 2.23 | 2.67 | 3.58 | 4.96 | 3.31 | 5.12 | 1.47 | 0.84 | 2.20 | 1.60 |
| 1893..... | 2.25 | **7.65** | 1.92 | **7.08** | 4.81 | 2.89 | 1.27 | 1.65 | 1.14 | 3.33 | 2.16 | 1.97 |
| 1894..... | 2.42 | 3.11 | 1.79 | 1.79 | 2.78 | **1.12** | 1.74 | 2.64 | 5.31 | 1.93 | 1.91 | 2.95 |
| 1895..... | 4.67 | 0.64 | 1.23 | 4.12 | 1.73 | 2.94 | 1.45 | 2.10 | 1.48 | 0.92 | 5.32 | 4.14 |
| 1896..... | 2.34 | 1.93 | 3.04 | 2.70 | 2.61 | 3.38 | **9.47** | 3.53 | 5.93 | 0.55 | 3.53 | 1.52 |
| 1897..... | 1.54 | 3.71 | 5.45 | 4.27 | 3.68 | 2.45 | 6.95 | 1.95 | 0.82 | 0.36 | **7.54** | 2.43 |
| 1898..... | 5.29 | 1.67 | 7.03 | 2.05 | 6.04 | 1.63 | 2.33 | **7.16** | 1.77 | 2.95 | 2.30 | 1.09 |
| 1899..... | 2.35 | 1.44 | 4.69 | 1.18 | 2.25 | 1.26 | 4.85 | 1.49 | 2.01 | 2.23 | 1.72 | 2.98 |
| 1900..... | 3.01 | 3.30 | 2.59 | 1.76 | 1.82 | 2.45 | 3.89 | 3.02 | 0.97 | 2.86 | 3.71 | **0.92** |
| 1901..... | 1.50 | 0.88 | 1.82 | 2.21 | 4.24 | 6.31 | 1.23 | 1.71 | 2.10 | 0.33 | 0.59 | 3.61 |
| 1902..... | 1.56 | 0.51 | 2.63 | 1.60 | **0.95** | **8.52** | 4.70 | 1.62 | 4.16 | 1.85 | 2.72 | 3.41 |
| 1903..... | 2.11 | 4.44 | 4.13 | 2.47 | 2.18 | 3.07 | 2.05 | 0.67 | 1.46 | 1.84 | 2.01 | 1.71 |
| 1904..... | 2.80 | 8.12 | 4.93 | 2.49 | 4.01 | 3.86 | 2.48 | 3.18 | 0.83 | 0.97 | **0.18** | 3.63 |
| 1905..... | **1.25** | 1.57 | 5.87 | 3.15 | 4,38 | 2.78 | 2.27 | 5.45 | 3.36 | 5.45 | 1.64 | 1.87 |
| 1906..... | 1.98 | 1.08 | 4.59 | 1.16 | 2.47 | 1.44 | 5.27 | 6.15 | 1.59 | 2.07 | 2.57 | 3.33 |
| 1907..... | 5.73 | **0.43** | 5.21 | 3.27 | 3.35 | 3.39 | 6.07 | 2.47 | 2.27 | 1.59 | 1.68 | 1.85 |
| 1908..... | 1.40 | 3.66 | 6.03 | 2.75 | 4.04 | 2.13 | 3.74 | 2.34 | **0.42** | 1.20 | 0.84 | 1.59 |
| 1909..... | 2.52 | 4.97 | 2.68 | 3.20 | 4.65 | 3.88 | 3.34 | 2.53 | 1.81 | 2.77 | 1.66 | 2.58 |
| 1910..... | 5.11 | 5.05 | **0.28** | 2.52 | 4.10 | 2.93 | 2.40 | **0.42** | 3.66 | 5.22 | 0.79 | 2.31 |
| 1911..... | 4.46 | 1.71 | 2.36 | 4.37 | 1.15 | 4.04 | 3.29 | 3.62 | 5.98 | 5.21 | 2.71 | 4.53 |
| 1912..... | 1.58 | 1.53 | 4.56 | 4.20 | 2.65 | 1.48 | 3.50 | 2.25 | 2.83 | 1.71 | 1.01 | 2.34 |
| 1913..... | **6.63** | 2.09 | **8.09** | 3.91 | 2.60 | 1.56 | 2.88 | 2.10 | 3.28 | 2.05 | 4.56 | 1.13 |
| 1914..... | 2.21 | 3.70 | 2.46 | 2.48 | 1.28 | 2.03 | 1.64 | 4.78 | 1.26 | 4.44 | 1.99 | 2.91 |
| 1915..... | 3.30 | 1.52 | 1.19 | 0.95 | 2.57 | 5.06 | 6.85 | 7.01 | 4.43 | 0.94 | 1.97 | 4.15 |
| 1916..... | 5.02 | 1.47 | 4.88 | 2.33 | 4.81 | 3.49 | **0.66** | 3.22 | 1.54 | 1.84 | 1.58 | 3.59 |

4. Study the frequency distributions of population with respect to age; Report of the Thirteenth Census, 1910, Chapter IV, Vol. 1, with special reference to the size of the various class intervals and note two general forms of stating the frequencies of the classes.

5. Examine the different forms of frequency distributions appearing in the report of the Medico-Actuarial Society's Investigations, Vols. I, II, III, IV; also in Biometrika, Agricultural Experiment Station Bulletins and in other accessible sources.

6. In which of the exercises of Chapter I is the data in the frequency distribution form?

**Plotting a Frequency Distribution.** The illustrative data at the beginning of this chapter is plotted by locating 14 equidistant points on a horizontal line, one for each height class from 61 inches to 74 inches inclusive. Then at the middle of each interval so obtained a vertical line is erected with a height proportional to the corresponding class frequency. In this way a point is obtained for each class.

As in Chapter II, a rectangle is constructed on each interval. It must be apparent that a rectangle in the case of the frequency distribution has in every case a significant statistical meaning—it is the frequency of the class. Hence the sum of the areas of all the rectangles is the total frequency of the distribution.

**Smoothing the Frequency Curve.** With the rectangles drawn, the smoothing of a frequency distribution is in no wise different from the smoothing of the data discussed in the preceding chapter. However, for the frequency curve the two rules of the *permanence* of *areas* have a stronger justification because of the more definite significance of the areas under the curve.

With practice in the construction of statistical diagrams and curves the rectangles may be dispensed with and the curve drawn by inspection, especially when the data contains a large element of uncertainty. Also the broken line obtained by joining the ends of the ordinates, called the **freuency polygon,** may be smoothed by inspection into the required curve.

### Exercises.

7. Smooth the illustrative data at the beginning of this chapter.

8. Smooth the frequency distribution of Chicago Top Beef Cattle Prices for 50-cent intervals.

9. Tabulate the same data to show the distribution for 25-cent classes.

10. Construct the smoothed frequency curve for the distribution of temperature and of precipitation at Columbus since 1879.

11. From data obtained from a financial paper construct the frequency distribution of the prices of preferred stocks for any one market day.

12. Do the same for a common stock.

13. Draw the smoothed curve of the following weight distribution of Ohio State University freshmen.

| Weight Class — | 102 | 107 | 112 | 117 | 122 | 127 | 132 | 137 | 142 | 147 | 152 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency — | 8 | 13 | 20 | 48 | 76 | 93 | 93 | 110 | 93 | 49 | 56 |

| Weight Class — | 157 | 162 | 167 | 172 | 177 | 182 | 187 |
|---|---|---|---|---|---|---|---|
| Frequency — | 31 | 22 | 13 | 11 | 3 | 2 | 9 |

The weight classes are here of width five pounds and the middle value of each class is taken as the class number. Class 187 includes all persons with weight greater than 184.

14. Construct the smooth curve of the distribution of ages of graduates from the Columbus Public Schools.

| Ages — | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|---|---|---|
| Numbers — | 0 | 7 | 45 | 186 | 114 | 61 | 8 | 0 |

13. Construct the frequency curve of the preferred stock data of Exercise 11.

14. Do the same for the common stock data of Exercise 12.

**Use of the Frequency Curve.** The frequency curve does not give a chronological picture of the variations in the data. Instead it shows the number of times that each value occurs. The frequency curves of precipitation for a dryer climate is located to the left of that for a more moist climate because months with small precipitation occur more frequently in the dryer region. The frequency curve of higher prices lies further to the right than does that of lower prices, so that by constructing the frequency curves it can be readily discovered which series of prices tends to be higher.

### Exercises.

15. Compare the Top Beef Cattle Prices of 1895 with those of 1915.

16. Compare the precipitation at Columbus with that of some other station.

**Typical or Representative Data.** Statistical data may be collected for the express purpose of exhibiting a chronological

or other statement of the variations. This sort of data is usually based on the complete enumeration of a given set of objects, as the census of population to apportion the members of the House of Representatives, or measures of stature for military purposes.

In discussing an increase in prices it is impossible to quote all prices; recourse must be had to a carefully selected list of prices. The condition of trade in certain industries is taken as indicative of the condition of all business. In comparing the prices of beef and the prices of corn the real object of investigation is to find an underlying connection between the two series of values—a connection which will hold good in any particular year. In such a study the historical statistics of the two price variations are in reality used as **representative,** as **typical,** of the manner in which the two prices are related. It is apparent that the frequency form of distribution is peculiarly adapted to typical data.

**The Errors in Representative Data.** The theory of enumerative statistics is simple in statement; the chief cares of the statistician are that all objects are counted and none counted more than once, and that an adequate and effective method of presentation is adopted. There are also complicated questions of the methods of collecting the data and of the limits of accuracy of the data but these are met with in data of either form.

Because it is practically impossible to secure **homogeneous** data; that is, data in which the values for all characteristics except those under consideration are the same for all variates, representative data must be examined for homogeneity. For instance, the persons whose heights are tabulated at the beginning of this chapter differ in age, early environment, physical condition, as well as in height so that the given distribution is in reality a distribution of a complex of attributes instead of merely the one attribute, height. Unless the influence of these various factors is carefully studied, serious errors may result from attempting to apply to another distribution the conclusions drawn from this distribution.

It may be shown that, from absolutely homogeneous material successive samples made up strictly at random, that is, without bias or prejudice, will most likely give materially differing

distributions. The extent of such errors must be understood in reasoning from one distribution to another.

Hence in working with typical or representative data care must be taken regarding (1) the limits of accuracy of the data; (2) the homogeneity of the data; (3) the errors of random sampling.

## CHAPTER IV.

**The Arithmetic Mean.** Let us add the January prices in the data of page 25, and then divide the sum by the number of items. The result is $8.33. In this way a number, **the arithmetic mean,** is obtained. The characteristic arithmetic property of this number is that each of the given data values may be replaced by it without altering the total sum of all the values.

It is usual to speak of the arithmetic mean simply as the **mean** unless, in order to distinguish the arithmetic mean from some other mean, there is special need for the defining word *"arithmetic."*

### Exercises.

1. Determine from the date of Exercise 1, Chapter I, the arithmetic mean of the monthly rainfall at Columbus, for March.

2. Determine from the data of Exercise 5, Chapter I, the arithmetic mean of the annual precipitation at Columbus.

3. Find from the data of page 25 the arithmetic mean of the 1895 Top Beef Cattle prices and compare with the 1915 mean.

4. On the assumption that the population of the United States increased uniformly from 1900 to 1910 find the value of the annual increase and then the estimated population for 1906.

5. Compute the arithmetic mean of the Monthly Top Beef Cattle prices for the years 1895 to 1916.

6. By first assigning each monthly price to the appropriate 50-cent class as on page 25 and computing the arithmetic mean of the prices when so altered determine the effect on the value of the arithmetic mean of substituting the class prices for the exact values. Use the class numbers in the computation and translate the result in terms of the proper interval.

7. In Exercise 5 there are 264 entries in the sum to be added. Show that much of the labor of the addition can be avoided by selecting the equal prices, then multiplying each by the number of times it occurs, and adding the resulting products to obtain the total sum of prices.

The results of Exercises 5, 6 and 7 suggest the computing of the mean from a frequency table in accordance with the following rule: *multiply each deviation by its frequency, add the resulting products, and divide this total sum by the total frequency.* The quotient is the value of the mean. Thus, from the frequency distribution of Top Beef Cattle Prices of Chapter III, obtained on page 26 — 6, 13, 35, 43, 30, 30, 18, 12, 15, 16, 17, 3, 6, 7, 1, — the mean price is given by the expression —

$$d = \frac{1 \times 6 + 2 \times 13 + 3 \times 35 + 4 \times 43 + 5 \times 30 + 6 \times 30 + 7 \times 18 + 8 \times 12 +}{252}$$

$$\frac{+ 9 \times 15 + 10 \times 16 + 11 \times 17 + 12 \times 8 + 13 \times 6 + 14 \times 7 + 15 \times 1}{252},$$

$= 6.00$, where d is the distance of the computed mean from the origin.

The mean class is thus 6.00; that is, 6 of the 50-cent intervals. This gives 7.25, the mid-value of class 6, as the mean price.

Whenever the frequency table is available the method just described is usually the shortest method for computing the value of the mean. However if the frequency distribution is not needed for any other purpose and especially if an adding machine is at hand the saving of time in the computation of the mean does not ordinarily justify the compilation of a frequency table merely for the one purpose of finding the mean.

The following is the computation for mean height from the data at the beginning of Chapter III.

Let us take the origin at height 60. Then the computation scheme will be as follows:

**Computation of the Mean.**

| *Class.* | *Deviation.* | *Frequency.* | *Dev. times Freq.* |
|---|---|---|---|
| 61 | 1 | 2 | 2 |
| 62 | 2 | 10 | 20 |
| 63 | 3 | 11 | 33 |
| 64 | 4 | 38 | 152 |
| 65 | 5 | 57 | 285 |
| 66 | 6 | 93 | 558 |
| 67 | 7 | 106 | 742 |
| 68 | 8 | 126 | 1,008 |
| 69 | 9 | 109 | 981 |
| 70 | 10 | 87 | 870 |
| 71 | 11 | 75 | 825 |
| 72 | 12 | 23 | 276 |
| 73 | 13 | 9 | 117 |
| 74 | 14 | 4 | 56 |
| | | 750 | 5,925 |

$$d = \frac{5,925}{750} = 7.9$$

TABLE II.

Hence the mean height is 7.9 classes; that is, 7.9 inches from the origin, and is therefore equal to 67.9 inches.

**Statistical Properties of the Arithmetic Mean.** What is the statistical significance and interpretation of the arithmetic mean? If a higher price were substituted for one of the January beef cattle prices the resulting arithmetic mean would be larger, but not so much larger as the individual price because in the process of obtaining the mean the price increase is divided by the total number of prices. Hence a *larger mean denotes that, as a whole the values of the distribution are greater, and a smaller arithmetic mean is to be interpreted as indicating a relative lower series of values.* And since all increases and decreases are to be divided by the number of varieties the changes in the value of the arithmetic mean are relatively smaller than those of the individual values. Thus a decrease of 50 cents must occur in each of the above prices in order to decrease the arithmetic mean by the same amount. A decrease of 50 cents in one-half the variates decreases the arithmetic mean by only 25 cents, and so on. That is, *the arithmetic mean is relatively more stable than is an individual measurement.* Thus if several groups of 750 students were measured for height and the frequency distribution tabulated and the means computed for each group it would be found that the means would differ but little while the frequency of any one class, 67 inches for instance, would vary considerably from distribution to distribution.

It is to be noted that a single increase of 50 cents in the price for one month has exactly the same effect on the value of the arithmetic mean as does a 10-cent increase in the prices of each of five months. But is this true statistically? Should the exceptionally high price be given so much weight? Should the person of exceptional height be emphasized so strongly in the group of persons whose height is measured?

That is, the value of the mean may not always be significant because a part of its value may be due to the presence of unduly large variates. Whether an item is unduly large can be determined only from a study of the data itself for the mean conveys no information whatever as to the distribution of the variates; it tells only of their general size. *That is, the statistical*

*function of the arithmetic mean is essentially to measure the size or magnitude of the data as a whole.*

**Theorem.** *In any distribution the sum of the deviations from the mean is zero.* That is, the sum of the positive deviations is equal to the sum of the negative deviations. The distance of the mean from any origin is obtained by taking the sum of the deviations from that origin and dividing by the total frequency, hence when this distance is zero the sum of the deviations must be zero.

**Weighted Arithmetic Mean.** An apparent modification of the arithmetic mean is illustrated by the following: It is desired to obtain an index of food prices by taking the mean of the price quotations of 15 articles of food. It is decided however, that one of the quotations should be given twice the weight of the other articles. This is done by multiplying this quotation by two and taking the double quotation in the total sum. The article is said to have a *weight* of two. The idea of weight introduces no new principles into the computation of the arithmetic mean.

**Adjustment or Graduation Formulas.** A class of adjustment formulas of wide and convenient adaptability to the smoothing of data are based on the arithmetic mean.

A series of terms not differing greatly from each other may be smoothed by replacing each by the mean of the five terms, for instance, of which the given term is the middle term. The distribution obtained by the first adjustment may in turn be similarly smoothed, and indeed the process may be repeated at pleasure. In this way the various graduation formulas of this type are built up. Next to the graphic method this is the simplest method for the smoothing of observations.

Extensive application of this method has been made in the graduating of mortality tables, and under the name of the **method** of the **moving average** it is often used in smoothing data in which the general trend is obscured by the presence of more or less regular fluctuations. In this case the number of classes grouped together should be determined by the lengths of the cycles of the fluctuations.* If the cycles are irregular in

* See King, Elements of Statistical Method, sec. 97. Also quarterly Publications of the American Statistical Society, Dec., 1915, and March, 1916.

length the method of the moving average is not likely to yield satisfactory results.

### Exercises.

8. Smooth the data of Table 1, by taking the means of each successive five terms, then of seven and finally of nine.

9. Apply the method of the moving average to smooth the data of the Top Beef Cattle Prices. Is the method highly applicable to this data?

10. Discuss the reliability of this method for terms at the end of the range.

11. Apply the five term method to the distribution of Ex. 5, Chap. I.

**The Geometric Mean.** Let the price of a certain article for each year from 1910 to 1915 be expressed as a percent of that of the preceding year as follows (assuming 100 for the 1910 price), 100, 105, 118, 109, 102, 115. The percent increase from 1910 to 1915 is obtained by multiplying together the five percents and is approximately 1.58. What uniform percent of increase will give the same percent of increase of 1915 over 1910? Let $(1 + r)$ be the constant multiplier or percent. Then we have

$$(1 + r)^5 = 105 \times 118 \times 109 \times 102 \times 115,$$
$$= 1.58415.$$
$$\text{and } (1 + r) = \sqrt[5]{1.58415},$$
$$= 1.096.$$

Each of the unequal increases in the series may therefore be replaced by the percent, 1.096, and still give the same product.

The population of continental United States in 1910 was 91,972,266; in 1900, 75,994,575. On the assumption of a uniform *rate* of increase during the decade what should be the value of this uniform rate in percent? As above, we have

$$(1 + r)^{10} = 91{,}972{,}266/75{,}994{,}575 = 1.21025.$$
$$\text{Hence } (1 + r) = \sqrt[10]{1.21025},$$
$$= 1.019.$$

It may be noted that according to this method the population in 1906 was equal to $75{,}994{,}575 \times (1.019)^6$.

The problem in the case of the arithmetic mean is to find a uniform number which, when substituted for each of

the variates, leaves the **total sum** unchanged. In problems similar to that just preceding it is a matter of finding a number which, when substituted for each of the given numbers, leaves the **product** of all the numbers unchanged; such a number is called the **geometric mean.**

### Exercises.

12. Compute the geometric mean of the following numbers: 2, 4, 8.

13. Compare from exercise 4, the 1906 population on the assumption of a uniform annual *increase* with that obtained from the assumption of a uniform annual *rate.*

For any but the simplest problems the computation of the geometric mean cannot be accomplished without the use of logarithms. The following computation of the geometric mean of student heights from the data of page 24 illustrates the process.

*The geometric mean height* $= (\cdot 61^{2} \cdot 62^{10} \cdot 63^{11} \cdot 64^{38} \cdot 65^{57} \cdot 66^{93} \cdot 67^{106} \cdot 68^{126} \cdot 69^{109} \cdot 70^{87} \cdot 71^{75} \cdot 72^{23} \cdot 73^{9} \cdot 74^{4})^{1/750}$

and 750 *log geo. mean* =

$$\begin{aligned} &2\ log\ 61 + 10\ log\ 62 + 11\ log\ 63 + \\ &38\ log\ 64 + 57\ log\ 65 + 93\ log\ 66 + \\ &106\ log\ 67 + 126\ log\ 68 + 109\ log\ 69 + \\ &87\ log\ 70 + 75\ log\ 71 + 23\ log\ 72 + \\ &9\ log\ 73 + 4\ log\ 74. \end{aligned}$$

Hence $log.\ geo.\ mean = \dfrac{1373 \cdot 70355}{750} = 1.83160$

and *geo. mean height* $= 67.86$.

### Exercises.

14. Compute the geometric mean of the Monthly Top Beef Cattle Prices.

15. Compete the geometric mean for the March precipitation at Columbus for the years since 1878.

**Properties of the geometric mean.** Unlike the arithmetic mean the geometric mean is most powerfully affected by the smaller deviations because a small factor in a product has a proportionately greater influence on the result of the multiplication than does a larger factor.

Each property of the arithmetic mean has a corresponding property for the geometric mean because the logarithm of the

geometric mean is the arithmetic mean of the logarithms of the deviations. From this logarithmic correspondence all the properties of the geometric mean can be derived* from those of the arithmetic mean. It is apparent, for instance, that the geometric mean applies to a series of deviations multiplied together in a way exactly parallel to that of the arithmetic mean and a series of terms to be added. Other parallels are, a chain of relative prices and a series of price increases; interpolation on the assumption of a uniform rate and of a uniform increase; compound interest and of simple interest.

**The Median.** Let the years 1879 to 1916 inclusive be arranged in order of the March precipitation beginning with the lowest. We then have with the data† measured to hundredths of an inch:

| | |
|---|---|
| 1910 | 0.28 |
| 1885 | 0.53 |
| 1889 | 0.66 |
| 1915 | 1.19 |
| 1895 | 1.23 |
| 1894 | 1.79 |
| 1901 | 1.82 |
| 1905 | 1.87 |
| 1893 | 1.92 |
| 1892 | 2.23 |
| 1911 | 2.36 |
| 1880 | 2.42 |
| 1914 | 2.46 |
| 1887 | 2.56 |
| 1900 | 2.59 |
| 1902 | 2.63 |
| 1909 | 2.68 |
| 1896 | 3.04 |
| 1883 | 3.20 |
| 1884 | 3.59 |
| 1879 | 3.77 |
| 1888 | 3.79 |
| 1886 | 3.90 |
| 1881 | 4.01 |
| 1903 | 4.13 |
| 1912 | 4.56 |
| 1906 | 4.59 |
| 1891 | 4.64 |
| 1899 | 4.69 |
| 1882 | 4.76 |
| 1904 | 4.93 |
| 1907 | 5.21 |
| 1897 | 5.45 |
| 1890 | 5.63 |
| 1908 | 6.03 |
| 1898 | 7.03 |
| 1913 | 8.09. |

* See Zizek, "Statistical Averages," Chapter III. Also Jevons, "On the Variation of Prices and the Value of the Currency since 1782," Jour. Roy. Stat. Soc., Vol. XXVIII, 1865. Galton, "The Geometric Mean" in Vital and Social Statistics," Proc. Roy, Soc., Vol. XXIX, 1897, p. 365. McAlister "The Law of the Geometrical Mean," the same, p. 367. Yule, "An Introduction to Statistics," p. 123.

† U. S. Weather Bureau Report, Columbus Station, 1917.

The middle year, 1883, in this ordered arrangement is called the **median** year with respect to March precipitation; the median precipitation of 3.30 inches, being that of the median year.

In general the **median** individual is defined as the individual so located that there are as many individuals with a greater value of the characteristic as with a less value; and the middle value of the measured characteristic is spoken of as the **median value** of the **characteristic.**

If the number of variates is even the medium is assumed to lie between the two middlemost variates.

It is obvious that the above median precipitation year might have been obtained by a simple process of counting and inspection of the data without the somewhat laborious process of arranging the variates in order.

### Exercises.

16. From the data of Exercise 2, Chapter III, determine the median Columbus monthly temperature, and the median year in respect to temperature.

17. From the price data of page 25 determine the median top beef cattle price.

18. From the data of Exercise 19, Chapter I, determine the median price for wheat.

When the data is in the form of a frequency distribution the computation of the position of the median is much facilitated. All that is necessary then is to start from one extremity of the distribution and include successive classes until half the total frequency is obtained. The only point of difficulty in this case is when the median is located within a class. Then it is necessary to interpolate within the median class for the more exact position of the median. To illustrate the method of interpolation let us find the median student height from the data at the beginning of Chapter III. Half of the number of variates is 375. Counting from the lower extremity we find, up to and including class 67, a frequency of 317, so that it is necessary to take 58 individuals from class 68. Hence we may assume that the position of the median will be 58/126 of a unit from the left boundary of class 68. Since this boundary is at 67.5 the median is located at 67.96 inches.

Geometrically, the median deviation locates the ordinate

which divides the area under the frequency curve into two equal parts.

### Exercises.

21. What is the median point of population as determined by the Bureau of the Census (see pp. 50-52, Vol. I., Report of the 13th Census)?

22. Distinguish the median point of population from the center of population.

**Quartiles.** Each half of the distribution, one on either side of the median, may be divided into two equal parts. These two points of division are the **First** and **Third Quartiles.**

The two quartiles and the median thus divide the variates into four classes of equal frequencies.

In data having predominately large frequencies near the center of the distribution the quartiles are relatively close to the median, and in widely scattered data the quartiles are relatively far from the median. This property of the quartiles is developed and applied in the next chapter.

The median can be found directly from the cumulative curve by drawing a horizontal line thru the point on the vertical scale corresponding to half the total frequency. The abscissa of the point of crossing of this horizontal line and the curve is the median deviation.

### Exercises.

19. By drawing the cumulative curve locate the median sudent height.

20. From the frequency distribution of top beef cattle prices of page 25 determine the median price by using the cumulative curve.

**Deciles.** The **decile variates** are the variates which separate the frequency into ten equal classes. The median is of course the fifth decile but the quartiles are not deciles. The chief use of the deciles, like that of the quartiles, is in determining the shape of the distribution.

### Exercises.

23. Determine the quartile precipitations from the data of Exercise 5, Chapter I.

24. Determine the decile precipitations from the data of Exercise 3, Chapter II.

25. Determine the quartile and the decile temperatures from the data of Exercise 2, Chapter III.

26. Determine the quartile prices from the top beef cattle prices of page 25.

27. Determine the quartile top beef cattle prices from the data in the form of a frequency distribution of the data of page 25.

In this problem the quartile prices must be obtained by a process of interpolation similar to that described for the median.

**Statistical Properties of the Median.** The value of the median ordinate depends not on the actual values of the variates but solely on the relative values. The data need be given with only enough exactness to permit the arrangement of the variates in order with respect to the attribute considered. Moreover, it is only the arrangement near the median value that must be carefully attended to, consequently the median can not give detailed information of the variates at the extremities of the ranges.

There is apparently *no apriori* reason why the value of the median should not show considerable variation from sample to sample taken from the same material, but in practice it is found that the median shows as high if not higher degree of stability than does the arithmetic mean. Thus if a second group of 750 students were measured as to height and the median computed it would most likely be found to differ only slightly from that of the group already discussed. This slowness of change in the median means that the median is not greatly affected by the presence of accidental and irrelevant influences. That is, differences in the value of the median are not likely to be merely accidental and hence the median measures significant properties of the material. For instance, a distribution of wages showing a higher median wage must be significantly a group of higher wages.

The properties just discussed together with the fact that the median can be located by the simple process of counting renders the median a highly important average in practical statistical work.

**The Probable Deviation.** The median variate divides the data into two classes of equal frequencies. Hence it is an even chance that an individual selected at random will fall into a designated one of the two classes. If the median height of freshmen students is 68 inches it is an even bet that a student concerning whose height nothing is known has a height less than 68 inches.

Likewise it is an even bet that a student selected at

random will have a height between the first and third quartiles. The range from the median to the third or first quartile, one-half of the range within which the chances are even for an individual measurement to lie, is called the **probable deviation.***

### Exercises.

28. Determine the probable deviation for top beef cattle prices.

29. Determine the probable deviation for monthly precipitation at Columbus; for monthly temperatures at the same station.

30. Show that the probable deviation is necessarily connected with the frequency distribution and not with a chronological distribution.

**The Mode.** Notice that, in the frequency distribution of student heights, class 68 has the greatest height and that the high point on the frequency curve is within the same class. The class of greatest frequency is called the **modal class** and the deviation with the highest ordinate the **modal deviation.** A **mode** is thus defined as a class or deviation of greatest frequency; more accurately, it is the class or deviation of greater frequency than that of either the class immediately greater or immediately less. This second definition allows for distributions having more than one mode.

### Exercises.

31. From the smoothed frequency curve of the data of page 27 determine the modal monthly precipitation.

32. Determine the modal March temperature for Columbus.

It is possible to locate the mode within a class by a process of interpolation similar to that described in the determination of the median but by far the easiest method is to construct the smooth frequency curve and determine the abscissa or deviation of the greatest ordinate.

When the data seems to have more than one mode care must be exercised in deciding whether to smooth out the apparent modes. In the frequency distribution of monthly temperatures it is evident that there are summer and winter modal temperatures. The telephone-calls data of Exercise 33 below shows more than one mode. On the other hand the data of age distribution reported by the United States Census Bureau

* Certain qualifications of this definition are discussed in Chapter V.

shows a tendency for the frequencies at the even ages to be larger than at the odd ages. This latter tendency is partly due to the fact that persons who are uncertain as to their exact age seem to show a preference for an even number. These apparent modes should be smoothed out. Data with essentially one mode is said to be **unimodal**; with more than one mode, **multimodal.**

## Exercises.

33. Smooth the following data of the telephone calls for one day at a business exchange* and locate the modes.

| Time .... | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 | 12-1 | 1-2 | 2-3 |
|---|---|---|---|---|---|---|---|---|---|
| Calls .... | 1595 | 3430 | 6389 | 6904 | 7282 | 7358 | 6361 | 5659 | 6186 |

| Time .... | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 |
|---|---|---|---|---|---|---|---|---|---|
| Calls .... | 6597 | 6510 | 6093 | 4508 | 4210 | 2289 | 1197 | 916 | 314 |

| Time .... | 11-12 |
|---|---|
| Calls .... | 12 |

34. Do the same for the following residence calls.**

| Time .... | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 | 12-1 | 1-2 | 2-3 |
|---|---|---|---|---|---|---|---|---|---|
| Calls .... | 1256 | 3796 | 6604 | 4098 | 4240 | 3816 | 5852 | 4421 | 3136 |

| Time .... | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 |
|---|---|---|---|---|---|---|---|---|---|
| Calls .... | 4344 | 3267 | 4541 | 4778 | 4039 | 2088 | 1176 | 655 | 187 |

35. Determine the modal classes for the top beef cattle prices.

**Statistical Properties of the Mode.** Because the necessary modifications are easily made for multimodal data the properties of the mode are here discussed only for a unimodal distribution.

Since the modal class or deviation is that of greatest frequency; that is, since more variates belong to that class than to any other, the mode is the most typical of all the variates of a distribution. If any one variate is to be selected as decriptive of the data the modal variate should be that variate. *The mode is accordingly said to define the type of the distribution.* The significance of the mode as a type depends, of

* By permission of Central Union Telephone Company, Columbus. Main Exchange.

** Same, North Exchange.

course, on the relative preponderance of its frequency. Thus the frequency of height 68 in the case of the student distribution of page 24 is 126 and the combined frequency of the classes near the modal class is a large percent of the total frequency. In the beef cattle prices of page 25 the modal class has a frequency of 43 and there is not as rapid falling off in the frequency on either side of this class as is shown by the height data. Hence in the price data the mode does not have as great significance as it does in the height data. Data showing a strong tendency to concentrate about the mode is said to be **highly stable** or **true to type.** Measures of trueness to type are discussed in the following chapter.

The position of the mode depends only on the values of a few variates so that the mode like the median gives little information of the extremes of the range.

The mode cannot be accurately determined by a simple process of arithmetic as can the median and the mean.

The mode being the predominating value, the type, the fashion, it is what is ordinarily in the popular mind when an average is spoken of. The statement that the average person spends one-third of his income in rent is most likely to mean that more persons spend about that per cent than any other per cent.

### Exercises.

36. Determine the modal class for each frequency distribution of Chapter III.

37. Show that the concept of mode does not apply to a curve of the historical type.

## CHAPTER V.

### THE FORM OF A DISTRIBUTION.

**Dispersion.** It is stated in the preceding chapter that the significance of the mode as a representative of the data depends on the extent to which the data conforms to the mode as a type. That is, if the sum of the frequencies near the mode is a relatively large per cent of the total frequency the modal deviation is highly typical and the data is not highly variable. The word variable is used because, if in the data a certain type does not predominate, different samples will have a tendency to show widely differing distributions. If, to illustrate, the modal frequency of a second distribution of the heights of 750 students is only 95 with a similar reduction in the other larger frequencies, this second distribution is not so true to the type expressed by the mode as is the first distribution.

To repeat, a distribution with small frequencies at the ends of the ranges and with the frequencies concentrated at a point is said to be **true to type**, to be **highly stable**. Let us investigate various methods of measuring the extent to which the data is **scattered** or **dispersed** about the class of concentration.

**Measures of Dispersion.** Because the breadth of the *range* depends on the usually uncertain data at the extremes it does not furnish a reliable measure of the extent to which the data is spread-out. As given on page 24 the range of student heights is 14 inches; the inclusion of a *single* student of height 58 inches would increase the range by more than twenty percent.

We have seen that in theory the dispersion should be measured from the mode but in practical statistical work the mean, median and mode differ so little in position that it is ordinarily permissible to measure the disperson from the mean.

The sum of the deviations about the mean is useless as a measure of dispersion because, as was proved on page 35, this sum is zero regardless of the spread or dispersion of the distribution.

**Mean Deviation.** Since the object in measuring disper-

sion is to determine the divergences of the variates from an average it is the amount of a devergence that counts and not its direction. Hence a logical measure of dispersion is obtained by adding the divergences, *all counted positive,* and then dividing the sum by the total frequency. This gives the **mean deviation.**

The form for the computation of the mean deviation is the same as for the arithmetic mean except that all deviations are measured from the mean, median or mode, whichever is chosen for the origin, and all negative signs are disregarded.

## Exercises.

1. Compute the mean deviation from the arithmetic mean of the Student Height Data of page 24.

Referring to the computation for the arithmetic mean on page 33, let us add a column obtained by taking the difference between the mean and each deviation and then multiply these differences by the respective frequencies and add the resulting products. This sum is then divided by the total frequency in order to obtain the mean deviation. We thus have:

### Computation of the Mean Deviation.

| *Class No.* | *Diff.* | *Freq.* | *Prod.* |
|---|---|---|---|
| 1 | 6.9 | 2 | 13.8 |
| 2 | 5.9 | 10 | 59.0 |
| 3 | 4.9 | 11 | 53.9 |
| 4 | 3.9 | 38 | 148.2 |
| 5 | 2.9 | 57 | 165.3 |
| 6 | 1.9 | 93 | 176.7 |
| 7 | 0.9 | 106 | 95.4 |
| 8 | 0.1 | 126 | 12.6 |
| 9 | 1.1 | 109 | 119.9 |
| 10 | 2.1 | 87 | 182.7 |
| 11 | 3.1 | 75 | 232.5 |
| 12 | 4.1 | 23 | 94.3 |
| 13 | 5.1 | 9 | 45.1 |
| 14 | 6.1 | 4 | 24.1 |
| | | 750 | 1,473.5 |
| | | | 1.9 |

Mean deviation = 1.9 classes.

Since each class interval is one inch the mean deviation is 1.9 inches.

TABLE III.

2. Compute the mean deviation about the arithmetic mean of the price data of page 25.

3. Compute the mean deviation about the median of the price data of page 25 and compare the result with that of Exercise 2.

4. Compute the mean deviation about the arithmetic mean of the precipitation data of Exercise 5, Chapter I, and of the temperature data of Exercise 6 of the same chapter.

5. From the frequency tables of Exercises 2 and 3 of Chapter III compute the mean deviation of monthly precipitation and of monthly temperature.

For purposes of comparing the stability of different distributions it is desirable to divide the mean deviation by the mean or median, whichever is used. When this is done the mean deviation is expressed as a fraction of the base average. For instance, it seems reasonable to say that a mean deviation of 0.3 with an arithmetic mean of 20 has the same significance as a mean deviation of 0.9 based on an arithmetic mean of 60.

### Exercises.

5. Compare the dispersions in Exercises 1, 2, 3, 4.

Because, as is presently proved, the mean deviation is least when taken about the median it is theoretically best to compute the mean deviation about that average. When so done there is a certain degree of standardization which is not attained with any other average as a base, but the point is not of great practical importance unless the median and the arithmetic mean differ markedly.

### Proof that the mean deviation is smallest when taken about the median.

Let P be a point on the line S-T between the points A and B. The sum of the deviations of P from A and B is, without regard to the sign of the negative deviation PA, PB + PA, and this sum is equal to AB. If P should lie without the segment AB the sum of the two deviations would be greater than AB. Likewise the sum of the distances of P from any other two points C and D is least when P lies between them. Hence the total sum of deviations of P from any number of points is least when there are as many points on one side of P as on the other; that is, when P is the median of the points.

S A C E P B D F T

### Exercises.

6. According to the measure supplied by the mean deviation which is the more variable, the monthly mean temperature or the monthly mean precipitation at Columbus?

7. From the data of heights on page 24 and the data of weights of Exercise 13, Chapter III, determine which is the more variable, student height or student weight.

**Statistical Properties of the Mean Deviation.** The mean deviation as a measure of dispersion has all the properties of a mean — it takes all the variates into account; it takes each variate according to its size and consequently may give more prominence to extreme variates than their statistical importance may warrant; it is computed by a simple process of arithmetic. Because in forming it only the numerical values of the deviations are used and all distinctions between positive and negative deviations are disregarded the mean deviation is not well adapted to certain statistical purposes for which the standard deviation, to be next discussed, is preeminently fitted.

Altogether the mean deviation is an index of dispersion of practical importance and should ordinarily be used either alone or in connection with other measures.

**The Standard Deviation.** The mathematically simplest device for eliminating negative signs is by squaring the terms. Hence if the difference between each deviation and the mean be squared, the sum of the squares added and the resulting sum divided by the total frequency the **mean squared deviation** thus obtained, is a measure of dispersion which is arithmetically more convenient than is the mean deviation.

The computation of the mean squared deviation differs from the computation of the mean deviation, which is illustrated under Exercise 1, only in that the deviation differences are squared before multiplication by the frequencies. It is of course possible to compute directly from the data without using the frequency table but only a slight error is introduced by the combining of the actual values into reasonably narrow classes and much labor is ordinarily saved because only one multiplication is then re-

quired for each *class* instead of for *each individual variate* as is necessary if the frequency distribution is not used.

### Exercises.

6. Determine the mean squared deviation about the arithmetic mean of the data of Student Heights.
7. Do the same for the Prices of Top Beef Cattle.
8. Do the same for Monthly Precipitation at the Columbus Station.
9. Do the same for Monthly Temperatures at the Columbus Station.

The above method of computing the mean squared deviation involves fractional differences in the deviations. By the following modification fractions can be avoided.

**Short Rule for the Mean Squared Deviation.** Select an integral deviation near the actual arithmetic mean and find the difference between each deviation and this selected deviation. Square each of the differences so obtained, multiply by the corresponding frequency, add and divide by the total frequency. The result is the mean squared deviation from the selected value. To obtain the mean squared deviation from the arithmetic mean all that is necessary is to subtract from the value just computed the square of the difference between the true arithmetic mean and the selected integral value. If the mean squared deviation about the actual arithmetic mean is denoted by the Greek letter $\sigma$, (sigma), and the mean squared deviation about any other point by the same symbol written with a prime, $\sigma'$; we have, on recalling that the letter $d$ is used to denote the deviation of the arithmetic mean from the origin, the following formula:

$$\sigma^2 = \sigma'^2 - d^2.$$

To prove this formula let the deviations from the original origin be denoted by X and the deviations from the arithmetic mean by $x$ and let the distance of the mean from the original origin be denoted by $d$. Then $X = x + d$ for each individual in the distribution and $x = X - d$.

The standard deviation is obtaineed by squaring each $x$ and dividing by the total frequency. Performing these operations we have

$$\sigma^2 = \frac{(X_1 - d)^2 + (X_2 - d)^2 + \dots}{N}$$

$$= \frac{X_1^2 + X_2^2 + \dots - 2d\,(X_1 + X_2 + \dots) - (d^2 + d^2 + \dots),}{N}$$

$$= \frac{N\sigma'^2}{N} - Nd^2/N = \sigma'^2 - d^2$$

Since $X_1 + X_2 + \dots$ is zero by the Theorem of page 35.

## Exercises.

10. Recompute by the shortened method the mean squared deviation about the arithmetic mean of the Student Heights.

Let us take the assumed origin at class 68. The mean squared deviation is then obtained by the following computation:

| *Class.* | *Dev.* | *Dev. squared.* | *Freq.* | *Prod.* |
|---|---|---|---|---|
| 1 | 7 | 49 | 2 | 98 |
| 2 | 6 | 36 | 10 | 360 |
| 3 | 5 | 25 | 11 | 275 |
| 4 | 4 | 16 | 38 | 608 |
| 5 | 3 | 9 | 57 | 513 |
| 6 | 2 | 4 | 93 | 372 |
| 7 | 1 | 1 | 106 | 106 |
| 8 | 0 | 0 | 126 | 0 |
| 9 | 1 | 1 | 109 | 109 |
| 10 | 2 | 4 | 87 | 348 |
| 11 | 3 | 9 | 75 | 675 |
| 12 | 4 | 16 | 23 | 368 |
| 13 | 5 | 25 | 9 | 225 |
| 14 | 6 | 36 | 4 | 144 |
| | | | 720 | )4032 |
| | | | | 5.60 |

$d = 68 - 67.9 = 0.1$; $d^2 = 0.01$.
Therefore $\sigma^2 = 5.60 - 0.01 = 5.59$
and $\sigma = 2.31$.

TABLE IV.

11. Using the shortened method compute the mean deviation for the Monthly Precipitation data.

**The Standard Deviation.** The square root of the mean squared deviation about the arithmetic mean is called the **Standard Deviation.**

From the formula $\sigma^2 = \sigma'^2 - d^2$ it is seen that any other mean squared deviation always exceeds the square of the standard deviation by the square of the difference between the assumed origin and the true arithmetic mean.

This gives a certain practical and theoretical preference to the standard deviation over that of any other mean squared deviation. For this reason, and because certain other computations are rendered simpler by so doing, the mean squared deviation about any other value than the arithmetic mean is seldom computed even tho the idea of trueness of type centers about the mode. Since the mean and the mode rarely differ by more than a small amount the square of this difference will be relatively still smaller and as a result, the difference between the square of the standard deviation and the mean squared deviation about the mode is ordinarily negligible.

**Properties of the Standard Deviation.** Since a small value for the standard deviation can arise only when the variates are closely concentrated about the mean or mode and since a large value must be due to a relatively high frequency of the variates near the extremes of the distribution, the standard deviation is a measure of the dispersion of the data. Because the effect of squaring is to diminish the importance of the smaller values and to exaggerate the importance of the larger values a small value for the standard deviation shows conclusively that the data is highly true to type and stable, while on the other hand a large value may to some extent be due to the presence of the larger frequencies of the extreme variates and hence not altogether significant. But even with this qualification the standard deviation is a thoroly practicable and reliable index of the dispersion of the data.

## Exercises.

12. Discuss the comparative variabilities of the distributions for which the standard deviations have been computed in the preceding exercises of this chapter.

13. Does a standard deviation of 2.4 for height denote a smaller variability than a standard deviation of 15 pounds for weight?

**The Coefficient of Variability.** As in the case of the mean deviation the significance of a value for the standard deviation depends on the size of the variates. A variation of 10 feet in a measurement of five miles is of the same degree of accuracy as a variation of 2 feet in one mile.

It is, therefore, reasonable to divide the standard deviation by the mean in order to express it as a fraction of the size of the variates. This quotient is ordinarily quite small, so that it is usual to multiply it by 100. The resulting coefficient — 100 times the standard deviation divided by the mean — is called the **coefficient of variability.**

### Exercises.

14. Compare the value of the coefficient of variability for height with that for weight as shown by the data of the preceding chapter.

15. Discuss the comparative variabilities of March and July temperatures as recorded at the Columbus Station.

16. Apply the coefficient of variability to determine which is the more variable, Columbus monthly temperature or Columbus monthly precipitation. Compare the variability of annual temperatures with that of precipitation.

17. Compare the value of the coefficient of variability with that of the quotient of the mean deviation by the arithmetic mean.

18. Discuss the comparative practical usefulness of the two indices of variability.

**The Quartiles as Measures of Dispersion.** The distance from the median to the third quartile is the interval that includes half the frequencies on the right of the median. The distance from the median to the first quartile is the interval that includes half the frequencies on the other side of the median. Now if these distances are relatively large it must mean that the frequencies at the center are not large in comparison with the total frequency. That is, if the first and the third quartiles are close together the distribution must be closely concentrated about the median; must be highly typical; must show a low degree of variability, because in every case one-half the total frequency is included between these two quartiles and if the interval is narrow the ordinates must be tall, that is the frequencies in the center must be predominatingly large, in order to include half the total frequency. If the data has a flat frequency curve so that the degree of variability is large

and the trueness to type small the two quartiles will be comparatively far apart.

Ordinarily the distance between the first quartile and the median is approximately equal to the distance from the median to the third quartile so that the distance from the median to the third quartile is taken as the index of dispersion of the distribution. This distance is called the **probable deviation.**

Since half the total number of frequencies are included between the two quartiles the chances are even that an individual of the group, selected at random, will have a deviation lying between the quartile deviations. In other words, the chances are even that an individual selected at random from the group will have a deviation numerically less than the probable deviation. If in one group of 750 students, for instance, it is an even bet that a student selected at random has a height between 66 and 70 inches and in a second group the range for even chances is from 67 to 69, the second group is said to be the *more true to type.*

**Formula for the Probable Deviation.** The probable deviation can always be found by the simple process of locating the quartiles. It is proved in the following chapter that for a certain special, tho very frequently occurring, form of distribution the probable deviation is equal to the standard deviation multiplied by a constant.

In symbols, we have $P.\ E. = 0.6745\ \sigma$, where the symbol $P.\ E.$ inherited from the theory of errors developed by *Gauss* denotes the probable deviation.

If the distribution is markedly unsymmetrical the above formula may not hold accurately and there are symmetrical distributions for which it does not hold exactly. But extreme accuracy in the matter of an index of dispersion is not necessary or desirable; the formula is generally used regardless of the form of the distribution.

### Exercises.

19. Compute the probable deviation of the Student Heights.
20. Compute the probable deviation of the Student Weights.
21. Compute the probable deviation of the Top Beef Cattle Prices.

**Probable Deviation of the Arithmetic Mean.** The arithmetic mean in the Student Height data of page 24 is 67.9 inches. The mean height of a second group of 750 students from the same student population would most likely not differ greatly from 67.9 but it is not at all likely that it would be exactly the same as that of the first group. Let group after group be taken and the value of the mean computed for each group. The values of these means would themselves form a frequency distribution from which a mean and probable deviation could be obtained.

Now if the student data is highly typical and stable the variation in the successive means will be within a small range and hence the probable deviation of the means will be relatively small. Let us assume the value which we have obtained by actual observation, namely 67.9, is the best estimate of the true mean of the height of all such students; that is, that the deviation of greatest frequency in the frequency distribution of means is 67.9. Then the probable deviation in this distribution will be the probable deviation of the mean. It can be proved that this probable deviation is obtained in accordance with the formula,

$$P.\ E.\ of\ mean = 0.6745\frac{\sigma}{\sqrt{N}}$$

**Exercises.**

22. Compute the probable deviation of the mean Monthly Precipitation for the Columbus Station.

23. Compute the probable deviation for the Monthly Top Beef Cattle Prices of page 25.

**Probable Deviation of the Standard Deviation.** The probable deviation of the standard deviation may be explained by a process of reasoning similar to that for the probable deviation of the mean. The formula for this probable deviation is:

$$P.\ E.\ of\ standard\ deviation = 0.6745\frac{\sigma}{\sqrt{2N}}$$

**Exercises.**

24. Compute the probable deviation of the standard deviation of Student Heights.

25. Compute the probable deviation of the standard deviation of Student Weights.

26. Which is the more variable, the standard deviation of Student Heights or of Weights?

**Statistical Significance of the Probable Deviation.** The statistical application of the probable deviation may be illustrated by the following questions: The mean height of a group of students is 67.9 with a probable deviation of 1.78 inches. The height of a student taken at random from a second group is 72 inches. What is to be concluded? That the two groups are taken from essentially the same populations or that they all are taken from distinctly different populations? That is, how many times may a deviation exceed the probable deviation and still be assumed to come from the same material? It must be apparent that this is a fundamental question in statistical analysis. Further discussion of it is deferred to the following chapter.

**The Deciles as Measures of Dispersion.** The position of the deciles shows the spread of the variates in the distribution. If the deciles near the middle of the distribution are close together and the deciles near the beginning and the end of the ranges are far apart the distribution is highly variable and not true to type. Because there are nine decile positions to observe in a distribution the decile is not so simple a measure of dispersion as is the quartile or standard deviation, tho this very fact of greater detail may in some cases be of advantage.

### Exercises.

27. By the use of the deciles compare the variability of monthly precipitation at the Columbus Station with that of monthly temperatures at the same station.

**Symmetrical and Asymmetrical Distributions.** The curve of student heights is essentially of the same shape to the right of the highest point as it is to the left. It is a symmetrical curve. (Fig. VI.) Statistically the fact of symmetry means in this case that there is no tendency for the students to be either tall or short; that there is no selection between the tall and the short; that the chances for a tall person to belong to the student group are equally as good as those of a short person; that there is absolutely no connection between being a member of this student group and being tall or being short.

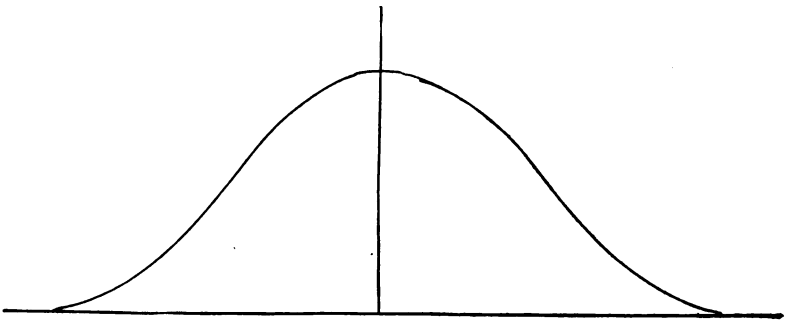

FIG. VI. A Symmetrical Curve.

On the other hand the curve of height of the members of a police force would have a longer range to the right than to the left because extremely short persons are excluded. The curve in this case is said to be **asymmetrical.** *Asymmetry in a curve denotes the presence of selection in the data;* of a dependence; of an expressed preference for certain values of the attribute.

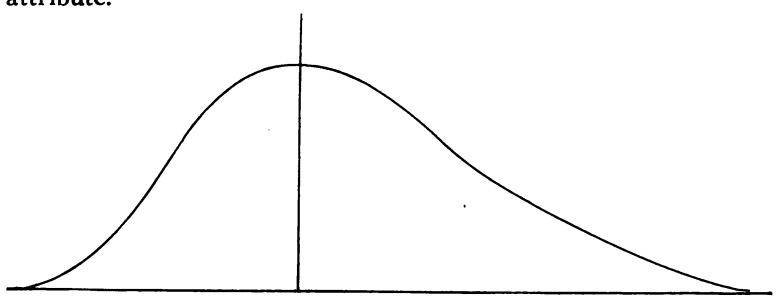

FIG. VII. An Asymmetrical or Skew Curve.

### Exercises.

28. Examine each frequency curve of Chapter III for symmetry and discuss the significance of each case of asymmetry.

**The Position of the Averages and Asymmetry.** In the symmetrical curve the mean, median and mode coincide. The cutting off of the range to the left tends to move the mean to the right because the longer deviations are to the right, and it has been seen that the mean is most affected by the longer or extreme deviations. This places the median at the left of the mean. The mode will tend to be moved to the left of

the median because both of the effect of the moving of the mean to the right and of the shortening of the left range with a consequent heaping up of the frequencies within the left half. The result is that the three averages are then in the order — mode, median, mean. It has been verified experimentally that for moderately asymmetrical distributions the distance of the median from the mode is about one-third the distance of the mean from the mode.

**Skewness.** An asymmetrical curve is said to be **skew. Skewness is positive** when the longer range is to the right and **negative** when the longer range is to the left.

**Measures of Skewness.** Since the mode and mean are separated to an extent depending on the degree of skewness present, a logical measure of skewness is the difference between the mean and the mode. Because a large difference between the positions of the mean and the mode in widely *spread-out* data may not be so significant as a smaller difference in highly *concentrated* data it is advisable to divide this difference by the standard deviation. Hence we have,

$$\text{Skewness} = \frac{\text{Mean} - \text{Mode}}{\sigma}.$$

## Exercises.

29. Compute the skewness of the following data of incomes:

Estimated Distribution of Income among the Single Women of Continental United States in 1890. (King, Wealth and Income, p. 224)

| Class | 0–$200 | $200–$300 | $300–$400 | $400–$500 | $500–$600 |
|---|---|---|---|---|---|
| No. in Thousands. | 10 | 70 | 560 | 530 | 280 |

| Class | $600–$700 | $700–$800 | $800–$900 | $900–$1000 |
|---|---|---|---|---|
| No. in Thousands. | 150 | 120 | 37 | 22 |

| Class | $1100–$1200 | $1200–$1300 | $1300–$1400 |
|---|---|---|---|
| No. in Thousands. | 12 | 8 | 5 |

30. Show that the above formula for skewness correctly indicates the sign of the skewness.

A **Second Measure of Skewness** is obtained as follows: Any measure of skewness must take into account the distinction between positive and negative deviations. The total sum of

deviations from the mean is zero regardless of the form of the distribution; the standard deviation involves the deviations as squares and hence obliterates the distinction between positive ang negative deviations. The mean cubed deviation, however, will serve as a measure of skewness. The longer deviations to the right, if the skewness is positive, will be more powerfully affected by the operation of cubing than will the shorter deviations to the left and hence the total sum of cubed deviations will be positive. It is well to extract the cube root of the mean cubed deviation and then in order to express the skewness as a fraction of the spread of the distribution to divide the result by the standard deviation.

### Exercises.

31. From the computation form of Exercise 1 compute, in accordance with the second method, the skewness for the student height distribution.

32. Do the same for the distribution of incomes.

## CHAPTER VI.

### THE NORMAL PROBABILITY CURVE.

**The Equation of a Frequency Curve.** As discussed in Chapter II, a smoothed curve is a graphic estimate of what would be the course of the data if it could be freed from accidental variations. The smoooth curve is therefore the geometric representation of a law of connection or variation. It shows, for instance, the variation of temperature with the seasons; the tendency for precipitation to depend on the month of the year; the most likely percent of students at each height.

The presence of an underlying law of connection in the data implies the presence of an algebraic law connecting the x and the y coördinates. The algebraic statement of the law expressing $y$ in terms of $x$ is called the **equation of the curve.**

If the equation is given, the ordinate can be computed for any abscissa and hence the curve can be located by plotting a sufficient number of computed points.

In some distributions it is possible to discover a law of connection directly from the data, and then without an extended computation to translate this law into the proper algebraic form. We shall discuss in this chapter the equation of only one type of curve — the normal curve. This form of curve is suited to the representation of a large class of distributions. And the theory of the normal curve can be made use of in the determination of the probable deviation and in the discussion of certain other properties even for a distribution to which it does not apply with sufficient accuracy to be adopted as the form of the smoothed curve.

**Statistical Theory of the Normal Curve.** The height of a person is the resultant sum of a large number of elements such as the length of certain bones, the widths of cartilages, the erectness of posture. And, in general, any statistical data can be analyzed into elemental components. *Whenever these elemental values are relatively small in comparison with the resultant values and at the same time each element is equally likely to take any value within a small range, then the resultant data is said to be* **normally distributed.**

With an absence of selection, as is assumed, it is reasonable to conclude that the resulting distribution will be symmetrical. And it is also apparent, after some consideration, that the frequencies at the center will be high and those at the ends of the range very small. It may be noted that in order to have a normal distribution it is not at all necessary that it be possible to actually compute the values of the elemental factors; it is only their existence under the above assumptions that is predicated.

**The Equation of the Normal Curve.** It can be mathematically demonstrated that the equation of the Normal Curve is,

$$y = \frac{N}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}$$

where $N$ is the total frequency of the distribution; $\sigma$, the standard deviation; $\pi$, the well known circle constant 3.14159; and e is a constant which is numerically equal to 2.71828. In this form of the equation $x$ is measured from the arithmetic mean as origin.

**The Graph of the Normal Equation.** Let us write the normal equation in the form, $y = \frac{N}{\sigma} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}$ and then in the form, $y = \frac{N}{\sigma} \cdot Z$ where $Z = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{x^2}{\sigma^2}}$

The tables of Sheppard give the value of $Z$ for each value of $x/\sigma$ from 0.00 to 6.00. Table V serves to illustrate the complete table.

**Table of Ordinates and Areas of the Normal Curve**

| $X/\sigma$ | $Z$ | *Areas.* | $X/\sigma$ | $Z$ | *Areas.* |
|---|---|---|---|---|---|
| 0.0 | 0.399 | 0.000 | 1.2 | 0.194 | 0.385 |
| 0.1 | 0.397 | 0.040 | 1.4 | 0.150 | 0.419 |
| 0.2 | 0.391 | 0.079 | 1.6 | 0.111 | 0.445 |
| 0.3 | 0.381 | 0.118 | 1.8 | 0.079 | 0.464 |
| 0.4 | 0.368 | 0.155 | 2.0 | 0.054 | 0.477 |
| 0.5 | 0.352 | 0.191 | 2.2 | 0.035 | 0.486 |
| 0.6 | 0.333 | 0.226 | 2.4 | 0.022 | 0.492 |
| 0.7 | 0.312 | 0.258 | 2.6 | 0.014 | 0.495 |
| 0.8 | 0.290 | 0.288 | 2.8 | 0.008 | 0.497 |
| 0.9 | 0.266 | 0.316 | 3.0 | 0.004 | 0.499 |
| 1.0 | 0.242 | 0.341 | 3.2 | 0.002 | 0.499 |

TABLE V.

With a table of values of $Z$ at hand the plotting of a normal curve is a simple matter of arithmetic. Each deviation from the mean is divided by the standard deviation and then the table is entered with the quotients, $x/\sigma$, and the values of $Z$ obtained by interpolation. Multiplication of the interpolated values by the ratio, $N/\sigma$, gives the successive values for $y$.

In the following illustrative plotting of the normal curve of the distribution of student heights the ordinates are computed for the boundaries (as the fractional deviations in the first column of Table VI denote) instead of the midpoints of the class intervals. This is done, as will be presently explained, in order to find the areas under the curve. In the computation scheme, the first column is for the deviations; the second for the deviations from the mean; the third, the deviations from the mean divided by the standard deviation; the fourth, the values of $z$ obtained from the table; and the fifth column shows the desired values for the ordinates. The sixth and seventh columns are explained on page 63.

**Table of Z's and Corresponding Areas for Student Height.**

| *Deviations.* | $X$ | $X/\sigma$ | $Z$ | $Y$ | *Areas.* | *Student Ht. Areas.* |
|---|---|---|---|---|---|---|
| 0.5 | —7.4 | —3.20 | 0.002 | 1. | 0.999 | 749 |
| 1.5 | —6.4 | —2.77 | 0.01 | 3.2 | 0.997 | 748 |
| 2.5 | —5.4 | —2.34 | 0.03 | 9.7 | 0.990 | 743 |
| 3.5 | —4.4 | —1.91 | 0.06 | 19.5 | 0.972 | 729 |
| 4.5 | —3.4 | —1.47 | 0.14 | 45.5 | 0.929 | 697 |
| 5.5 | —2.4 | —1.04 | 0.23 | 74.7 | 0.851 | 638 |
| 6.5 | —1.4 | —0.61 | 0.33 | 107.1 | 0.729 | 547 |
| 7.5 | —0.4 | —0.17 | 0.39 | 126.6 | 0.567 | 425 |
| 7.9 | —0.0 | —0.00 | 0.40 | 120.0 | 0.500 | 375 |
| 8.5 | +0.6 | +0.26 | 0.39 | 126.6 | 0.602 | 452 |
| 9.5 | +1.6 | +0.69 | 0.31 | 100.7 | 0.745 | 559 |
| 10.5 | +2.6 | +1.13 | 0.21 | 68.2 | 0.871 | 653 |
| 11.5 | +3.6 | +1.56 | 0.12 | 39.0 | 0.941 | 706 |
| 12.5 | +4.5 | +1.99 | 0.06 | 19.5 | 0.997 | 733 |
| 13.5 | +5.6 | +2.43 | 0.02 | 6.5 | 0.993 | 745 |
| 14.5 | +6.6 | +2.86 | 0.01 | 3.2 | 0.998 | 749 |

TABLE VI.

The computed points are now plotted and the curve drawn.

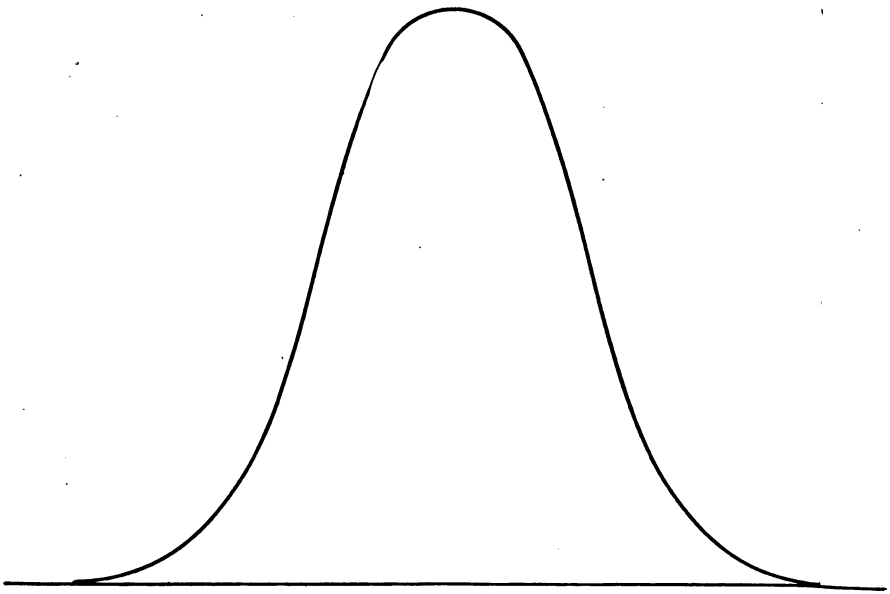

FIG. VIII. The Normal Curve of Student Heights.

In Figure VIII the characteristic bell shape of the normal curve is seen. The ordinates at the center do not change rapidly. As the deviations increase the ordinates first decrease rapidly and then more slowly until the curve flattens out so as to be almost coincident with the horizontal axis.

It is mathematically evident from the form of the equation of the normal curve that in a distribution with a *large standard deviation* the values for $y$ are relatively small near the center and relatively large for the greater deviations. That is a large value for $\sigma$ indicates a flat normal curve.

## Exercises.

1. Plot a normal curve for the distribution of weight according to the data of page 29.

2. Compare the curve obtained in Exercise 1 with the smooth curve of Chapter III. How closely do they coincide?

**Areas under the Normal Curve.** For each class the area under the frequency curve and between the class limits gives the smoothed class frequencies. In Sheppard's Table* the areas are given for each value of $x/\sigma$ from 0.00 to 6.00. Table V is sufficiently complete for illustrative purposes. The tables give the areas from $x/\sigma = 0$ to each designated value of $x/\sigma$. (In Sheppard's Tables 0.50,000,000 is added to these areas). Hence to determine the class areas from the tabular values appropriate subtractions are necessary, except in the case of the central area which is obtained by adding the two central fractional areas. It must be remembered that the actual area is finally obtained by multiplying the values just computed by the total frequency, and not by the total frequency divided by the standard deviation.

The sixth column of Table VI contains, for the frequency distribution of student heights, the areas from the origin; the class areas, obtained by subtraction or addition, are entered in the second column of Table VII. For comparison the original frequencies are placed in the third column of the same table.

**The Adjusted Distribution of Student Heights.**

| *Class.* | *Areas.* | *Orig. Freq.* | *Pos. Diff.* | *Neg Diff.* |
|---|---|---|---|---|
| 1 | 1 | 2 | .. | 1 |
| 2 | 5 | 10 | .. | 5 |
| 3 | 14 | 11 | 3 | .. |
| 4 | 32 | 38 | .. | 6 |
| 5 | 59 | 57 | 2 | .. |
| 6 | 91 | 93 | .. | 2 |
| 7 | 122 | 106 | 16 | .. |
| 8 | 127 | 126 | 1 | .. |
| 9 | 107 | 109 | .. | 2 |
| 10 | 94 | 87 | 7 | .. |
| 11 | 53 | 75 | .. | 22 |
| 12 | 27 | 23 | 4 | .. |
| 13 | 12 | 9 | 3 | .. |
| 14 | 4 | 4 | .. | .. |
| | 748 | 750 | 36 | 38 |

TABLE VII.

* "Tables for Statisticians and Brometricians." Table II.

The goodness of fit of this normal curve is indicated by the differences of the fourth and fifth columns. The differences are taken positive when the adjusted values exceed the original frequencies. The sum of the positive and of the negative differences shows a fairly close fit, though the size of the individual differences must also be taken into account in estimating the closeness of fit.

### Exercises.

3. Test the closeness of fit of the normal curve of student weights plotted in Exercise 1.

4. Compare the closeness of fit of the normal curves of weight and height.

**Preliminary Determination of Normality.** Before fitting a curve to a given distribution the data should be analyzed to determine whether the fundamental conditions of normality — whether each value of the data is the sum of a large number of relatively small elements each one of which is as likely to have one value as another, etc. — is satisfied. The data should be plotted and the smooth curve drawn by the methods of Chapter II. If the one or the other or both of these tests seem to indicate a normal distribution a normal curve should be fitted.

A more elaborate test for normality is to compute the mean cubed and the mean fourth power deviation. Unless the mean cubed deviation is very small the distribution possesses too much skewness or asymmetry to be closely fitted by a normal curve. It can be shown that for a distribution to be normal the mean fourth power of the deviations divided by the square of the mean squared deviation must be equal to 3. The variation that may be allowed from these two arithmetical tests is shown by Tables XXXVII and XXXVIII of *"Tables for Statisticians and Biometricians."*

However, a practically conclusive test of the appropriateness of the normal curve is that of comparing the adjusted with the original frequencies.

### Exercises.

5. Discuss the advisability of attempting to fit a normal curve to the precipitation data of page 7.

6. Is it likely that the frequency distribution of March temperatures will be more nearly normal than the distribution of temperatures for all months?

7. Discuss the probable fit of a normal curve in the case of the top beef cattle prices.

8. Is it likely that a normal curve will fit the income data of page 57.

9. What does a divergence from normality indicate?

10. What reasons are there for thinking that the distribution of grades in a large class of students should be normal?

11. What reasons might exist for thinking that data of prices might not be normally distributed?

**Probable Deviation in a Normal Distribution.** The quartiles divide the two halves of the area into equal parts; hence, in Table V, the value of $x/\sigma$ which corresponds to an area of 0.25, gives the value of the probable deviation. This value of $x/\sigma$ is there found equal to 0.6745. Therefore the deviation of the quartile is 0.6745 times the standard deviation. This demonstrates the rule for obtaining the probable deviation—multiplying the standard deviation by 0.6745.

The formulas for the probable deviation of the arithmetic mean and of the standard deviation referred to in the preceding chapter are derived on the assumption that the two constants are each normally distributed.

It can be shown mathematically that even when the form of distribution is distinctly non-normal the ordinary rules for finding the probable deviations hold with an approximation close enough for practical purposes, and experimentation with different forms of distributions bears out the mathematical conclusions.

### Exercises.

12. What is the deviation corresponding to the ordinate which marks off three-fourths of the area to the right of the center?

13. What part of the area under the normal curve is included between the median and the ordinate with a deviation of two times the standard deviation? Three times the standard deviation? Four times the standard deviation?

The results of Exercise 13 show that *the occurrence of a deviation of three times the standard deviation is highly improbable.* That is, a deviation greater than about three times the standard deviation must significantly indicate that the measurement is not that of an individual taken from the same material; it does not belong to the same distribution but to another dis-

tribution which has some conditions different from the first. To illustrate, the standard deviation of student heights is 2.36 inches and the mean height is 67.9 inches. One would according to this theory be justified in concluding that a person with a height of 75 inches ($67.9 + 3 \times 2.36 = 74.98$) does not belong to the student group.

13. Does the theory of Exercise 13 accord with the actual distribution of the student heights?

14. Does the theory of Exercise 13 accord with the actual distribution of the student weights? With the distribution of monthly precipitation for the Columbus Station? With the distribution of the monthly temperatures for the Columbus Station?

15. On the basis of the results of Exercises 13 and 14 and on other investigations of a similar nature discuss the practical applicability of the present theory of the probable deviation.

While it is not advisable to place implicit confidence in the tests furnished by the theory of probable deviations to the extent that the results which it indicates are accepted without some independent verification, or at least justification, yet when used with judgment it is an extremely valuable aid in practical statistical work. In every case it establishes cautionary limits, as, for instance, one would not ordinarily be justified in concluding that a variate with a deviation much greater than two or three times the standard deviation belonged to the distribution. On the other hand if a number of measurements of height should each consistently exceed those of the student distribution it might then be concluded with much certainty that the individuals measured were taken from a population distinctly different from the student population. And the conclusion would be justified even tho the deviations were considerably less than two or three times the standard deviation.

## CHAPTER VII.

### THE CORRELATION TABLE.

From the records of the physical measurements a tabulation was made of the heights of the students whose weight was from 130 to 134 pounds—a weight class which may be denoted by the middle weight, 132 pounds — and the following distribution obtained:

| Height | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number | 2 | 5 | 4 | 19 | 18 | 18 | 17 | 8 | 8 | 4 | 3 | 1 | 1 |

The distributions were likewise obtained for each other five-pound interval from 102 to 187 pounds. Instead of writing each of these distributions separately it is more convenient to write them together in one table called, for reasons explained on page 73, a **correlation table.** In this way we have Table VIII.

**Correlation Table of Height and Weight.**

HEIGHT IN INCHES.

| WEIGHT IN POUNDS. | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | To'ls |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 102 | .. | .. | 1 | .. | .. | 2 | 3 | 1 | 1 | .. | .. | .. | .. | .. | 8 |
| 107 | .. | 3 | 1 | 5 | 2 | 1 | 1 | .. | .. | .. | .. | .. | .. | .. | 13 |
| 102 | .. | .. | 2 | 7 | 3 | 3 | 3 | .. | .. | .. | 2 | .. | .. | .. | 20 |
| 117 | .. | 2 | 2 | 10 | 9 | 6 | 6 | 7 | 2 | 2 | 2 | .. | .. | .. | 48 |
| 122 | .. | 1 | 4 | 2 | 12 | 17 | 16 | 14 | 4 | 5 | .. | .. | 1 | .. | 76 |
| 127 | 1 | 1 | 1 | 7 | 7 | 11 | 15 | 16 | 18 | 9 | 5 | 2 | .. | .. | 93 |
| 132 | .. | 2 | .. | 4 | 9 | 18 | 18 | 17 | 8 | 8 | 4 | 3 | 1 | 1 | 93 |
| 137 | .. | 1 | .. | 3 | 4 | 14 | 20 | 24 | 21 | 11 | 9 | 2 | .. | 1 | 110 |
| 142 | .. | .. | .. | .. | 7 | 12 | 10 | 17 | 17 | 8 | 15 | 5 | 2 | .. | 95 |
| 147 | .. | .. | .. | .. | 2 | 3 | 7 | 5 | 12 | 9 | 8 | 3 | .. | .. | 49 |
| 152 | .. | .. | .. | .. | 2 | 2 | 3 | 14 | 10 | 12 | 11 | .. | 1 | 1 | 56 |
| 157 | .. | .. | .. | .. | .. | 4 | 1 | 6 | 7 | 5 | 7 | 1 | .. | .. | 31 |
| 162 | 1 | .. | .. | .. | .. | .. | 2 | 2 | 3 | 8 | 2 | 2 | 2 | .. | 22 |
| 167 | .. | .. | .. | .. | .. | .. | .. | 1 | 2 | 6 | 1 | 2 | 1 | .. | 13 |
| 172 | .. | .. | .. | .. | .. | .. | .. | 1 | 1 | 1 | 6 | 2 | .. | .. | 11 |
| 177 | .. | .. | .. | .. | .. | .. | .. | .. | .. | 1 | 1 | .. | .. | 1 | 3 |
| 182 | .. | .. | .. | .. | .. | .. | 1 | .. | .. | .. | .. | .. | 1 | .. | 2 |
| 187 | .. | .. | .. | .. | .. | .. | .. | 1 | 3 | 2 | 2 | 1 | .. | .. | 9 |
| Totals | 2 | 10 | 11 | 38 | 58 | 93 | 106 | 126 | 109 | 87 | 75 | 23 | 9 | 4 | 750 |

TABLE VIII.

The writing of the distributions in this compact tabular form greatly facilitates the study and comparison of the different distributions.

### Exercises.

1. Notice that there is a decided increase in weight with an increase in height; that there are no extremely tall persons in the group who are at the same time extremely light in weight; that there are practically no persons who are both short and extremely heavy.

2. Note that there is a closer connection between height and weight for the shorter and lighter individuals than for persons with medium values of the two characteristics.

**The Construction of a Correlation Table.** Let us construct the correlation table of monthly precipitation and monthly mean temperatures for the Columbus Station. The data is given under Exercises 2 and 3 of Chapter III. Let the horizontal scale refer to temperatures and let each class of this scale have a width of five degrees. The vertical scale will then refer to precipitation and let the width of classes be taken as one-half inch. The scales are written across the top and down the left hand margin respectively in order to leave room for the summations across the bottom and down the right hand margin. Under this arrangement of the scales $y$ increases in value from top to bottom and hence the positive direction for $y$ is downward.

In constructing the table it is convenient to rewrite the data according to classes and at the same time to combine the two distributions. There is no need for retaining the dates but care must be taken that the measures from exactly the same months are written together. This is done by starting with January, 1879, and proceeding with the Januarys and then February, 1879, and so on in order. The temperature figures are written first in each pair of numbers, and the lower limit is written as the class number of each class. Thus $\begin{matrix}25\\1.5\end{matrix}$ refers to a month with a mean temperature from 25 to 29 degrees inclusive and with a precipitation from 1.5 to 1.9 inches inclusive. In this way there will be built up a table of the following form:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 25 | 40 | 20 | 30 | 25 | 20 | 20 | 20 | 25 | 25 |
| 1.5 | 4.0 | 2.0 | 4.5 | 3.0 | 2.0 | 3.5 | 4.0 | 2.0 | 3.5, etc. |

Next the rulings must be made for the table. The tabulation proceeds in the following manner: for the first pair of numbers find the 25 column and drop down this column to the precipitation class 1.5 and mark a score; then to the 40 column and down to the 4.0 class and tally; then to column 20 and precipitation class 2.0; etc.

The diagram of tallies, usually dots, is called the **Scatter Diagram.**

**Correlation Table of Temperature and Precipitation.**

PRECIPITATION IN INCHES.

| TEMPERATURE IN DEGREES. | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 65 | 70 | 75 | To'ls |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.5 | .. | 3 | .. | 2 | 2 | 1 | 1 | 4 | 1 | 1 | 1 | 3 | 1 | 20 |
| 1.0 | .. | 2 | 4 | 5 | 2 | 2 | .. | 3 | 3 | 3 | 4 | 4 | 3 | 35 |
| 1.5 | 1 | 2 | 4 | 7 | 8 | 5 | 4 | 6 | 6 | 4 | 5 | 10 | 4 | 66 |
| 2.0 | 2 | 4 | 5 | 5 | 6 | 6 | 6 | 5 | 3 | 3 | 6 | 7 | 7 | 65 |
| 2.5 | .. | 1 | 3 | 5 | 5 | 3 | 3 | 7 | 6 | 5 | 5 | 16 | 2 | 61 |
| 3.0 | .. | 1 | 5 | 7 | 2 | 4 | .. | 5 | 3 | 1 | 5 | 5 | 5 | 43 |
| 3.5 | .. | 1 | 5 | 2 | 2 | 6 | 2 | 2 | 1 | 4 | 3 | 4 | 6 | 38 |
| 4.0 | .. | 1 | 2 | 2 | 5 | .. | 2 | 4 | 2 | 4 | 2 | 5 | 1 | 30 |
| 4.5 | .. | 1 | .. | 6 | 3 | 3 | 1 | 2 | 1 | 2 | .. | 3 | 2 | 24 |
| 5.0 | .. | .. | 2 | .. | 4 | 4 | 1 | 2 | 1 | 1 | 5 | 6 | 1 | 27 |
| 5.5 | .. | .. | .. | .. | 3 | 1 | .. | 1 | .. | 2 | 1 | 3 | .. | 11 |
| 6.0 | .. | .. | .. | 1 | .. | 2 | .. | .. | 1 | 1 | .. | 3 | 1 | 9 |
| 6.5 | .. | .. | .. | .. | 2 | .. | .. | .. | .. | .. | .. | 1 | .. | 3 |
| 7.0 | .. | .. | .. | .. | .. | .. | 1 | 1 | .. | 1 | 1 | 1 | 2 | 7 |
| 7.5 | .. | .. | .. | 1 | .. | 1 | .. | .. | .. | 1 | .. | .. | .. | 3 |
| 8.0 | .. | .. | .. | .. | .. | 1 | .. | .. | .. | .. | .. | .. | .. | 1 |
| 8.5 | .. | .. | .. | .. | .. | .. | .. | .. | .. | 1 | 1 | .. | .. | 2 |
| 9.0 | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | 0 |
| 9.5 | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | .. | 1 | .. | 1 |
| Totals | 3 | 16 | 31 | 43 | 44 | 42 | 21 | 44 | 30 | 35 | 39 | 73 | 35 | 456 |

TABLE IX.

Table IX, the correlation table, is made from the Scatter Diagram by inserting the frequencies in the place of the tallies.

## Exercises.

3. Do wet months uniformly occur with warm months? or is there more of a tendency for wet and cold or cool months to be associated?

4. What may be said as to the tendency for dry and warm months to be associated? for dry and cool months?

5. Does there seem to be as close a connection between precipitation and temperature as between height and weight?

6. Is it not possible that the real connection between precipitation and temperature in this table is obscured by the fact that data for all four seasons is thrown together? Explain.

**Definitions and Symbols.** The properties, as height and weight or temperatures and precipitation are called the **attributes** or **characteristics.**

The horizontal deviations are called the **x classes** or **deviations,** and the vertical, the **y classes** or **deviations.** Each subclass or subgroup thus has a value of $x$ and of $y$ associated with it. It is convenient to number the $x$ and $y$ classes from left to right and from top to bottom, respectively, and use these numbers for class numbers instead of the actual class values. Thus there are 17 persons with height 66 inches and weight 122 pounds; and 4 months with a mean temperature of from 40 to 45 degrees and a precipitation of from 3.0 to 3.5 inches. In terms of $x$ and $y$, the subclass $x = 6$, $y = 5$ has a frequency of 17; the subclass $x = 5$, $y = 6$ has a frequency of 4 months.

The columns and rows are spoken of as **arrays**; the columns as **y-arrays of type x** and the rows as **x-arrays of type y.** Or the concrete names of the data may be given to the arrays — the weight array of height 67 inches; the height array of weight 132 pounds; the precipitation array of temperature 40 degrees, etc. It should be noted that the weight array of height type 67 inches is the distribution with respect to weight of the persons having a height of 67 inches; the precipitation array of type 40 degrees is the precipitation distribution of the months having a mean temperature of 40 degrees.

A $y$ array of type $x$ and an $x$ array of type $y$ are said to be arrays of **opposite sense.** Two $y$ arrays or two $x$ arrays are arrays of the **same sense.**

The frequency of a $y$ array is denoted by the symbol $n_x$ where $x$ is the type of the array. The frequency of an $x$ array is denoted by the symbol $n_y$, where $y$ is the type. The frequency of a subclass is denoted by the symbol $n_{xy}$, where $x$ and $y$ are the deviations of the subclass; that is, the types of its two arrays. Thus, $n_{61} = 2$; $n_{132} = 93$; $n_{66 \cdot 142} = 12$, or if the simpler class numbers are used, $n_{:1} = 2$; $n_{:7} = 93$; $n_{6:9} = 12$. When the lat-

ter form of class numbers is employed it is necessary to distinguish, between $x$ and $y$ class numbers by means of a colon. Sometimes the distinction between $x$ and $y$ deviations or class numbers is made by the use of subscripts as $n_{x_1\, y_2}$.

### Exercises.

7. Write the values of $n_x$ for $x = 2, 4, 9$ in the precipitation data.

8. Write the values of $n_{2:4}$ $n_{9:2}$ for both the height-weight and the precipitation-temperature data.

9. Practice stating the frequencies of the various arrays and subgroups; e. g. the frequency of the weight array of type 8 (68) is 126.

10. Note that $n_{1:7} + n_{2:7} + n_{3:7} + \ldots\ldots n_{14:7} = n_{:7} = 93$, for the height-weight data.

11. Write other statements in the form of that of Exercise 10.

The mean of the vertical column of totals is called the mean of all the weights, and in general, the **mean of all the y's**; and is denoted by the symbol $\bar{y}$. It is the mean of the vertical deviations of the variates when unclassified with respect to the horizontal attribute; the mean weight for all heights; the mean monthly precipitation disregarding temperature; the mean monthly precipitation for all temperatures taken together.

Likewise, the **mean of all the x's** is denoted by the symbol $\bar{x}$.

The means of the weight arrays are denoted by the symbols, $\bar{y}_{61}$, $\bar{y}_{62}$, $\bar{y}_{63}$. In general the **mean of the y-array of type x** is denoted by the symbols $\bar{y}_x$. The **mean of the x-array of type y** is denoted by the symbol $\bar{x}_y$.

### Exercises.

12. From the following data construct the correlation table of top hog and top beef cattle prices at Chicago.

## Chicago Monthly Top Hog Prices.

| Years. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1916. | $ 8.10 | $ 8.90 | $10.10 | $10.05 | $10.35 | $10.15 | $10.25 | $11.55 | $11.60 | $10.35 | $10.35 | $10.80 |
| 1915 | 7.40 | 7.25 | 7.05 | 7.90 | 7.95 | 7.95 | 8.12 | 8.05 | 8:50 | 8.95 | 7.75 | 7.10 |
| 1914 | 8.00 | 8.90 | 9.00 | 8.95 | 8.67 | 8.50 | 9.30 | 10.20 | 9.75 | 9.00 | 8.25 | 7.75 |
| 1913 | 7.80 | 8.70 | 9.62 | 9.70 | 8.85 | 9.00 | 9.62 | 9.40 | 9.65 | 9.10 | 8.30 | 8.15 |
| 1912 | 6.70 | 6.57 | 7.95 | 8.20 | 8.05 | 7.30 | 8.50 | 9.00 | 9.27 | 9.42 | 8.30 | 7.85 |
| 1911 | 8.30 | 7.90 | 7.35 | 6.90 | 6.50 | 6.72 | 7.55 | 7.95 | 7.80 | 6.90 | 6.72 | 6.60 |
| 1910 | 9.05 | 10.00 | 11.20 | 11.00 | 9.35 | 9.80 | 9.60 | 9.70 | 10.10 | 9.65 | 8.70 | 8.10 |
| 1909 | 6.70 | 6.95 | 7.15 | 7.60 | 7.55 | 8.20 | 8.45 | 8.32 | 8.60 | 8.40 | 8.45 | 8.75 |
| 1908. | 4.72 | 4.70 | 6.35 | 6.45 | 5.90 | 6.67 | 7.10 | 7.10 | 7.60 | 7.20 | 6.40 | 6.15 |
| 1907. | 7.05 | 7.25 | 7.10 | 6.90 | 6.65 | 6.42 | 6.65 | 6.72 | 7.00 | 7.00 | 6.32 | 5.30 |
| 1906. | 5.72 | 6.42 | 6.55 | 6.82 | 6.67 | 6.85 | 7.00 | 6.75 | 6.82 | 6.85 | 6.50 | 6.55 |
| 1905. | 5.00 | 5.12 | 5.55 | 5.72 | 5.65 | 5.70 | 6.17 | 6.45 | 6.20 | 5.80 | 5.25 | 5.35 |
| 1904 | 5.20 | 5.30 | 5.82 | 5.50 | 4.95 | 5.40 | 5.90 | 5.80 | 6.37 | 6.30 | 5.25 | 4.87 |
| 1903. | 7.10 | 7.65 | 7.87 | 7.65 | 7.15 | 6.45 | 6.10 | 6.20 | 6.45 | 6.50 | 5.50 | 4.90 |
| 1902 | 6.85 | 6.60 | 6.95 | 7.50 | 7.50 | 7.95 | 8.25 | 7.95 | 8.20 | 7.92 | 6.95 | 6.80 |
| 1901 | 5.47 | 5.65 | 6.20 | 6.25 | 6.05 | 6.30 | 6.40 | 6.75 | 7.37 | 7.10 | 6.30 | 6.90 |
| 1900 | 4.92 | 5.10 | 5.55 | 5.85 | 5.57 | 5.42 | 5.55 | 5.57 | 5.70 | 5.55 | 5.12 | 5.10 |
| 1899 | 4.05 | 4.05 | 4.00 | 4.15 | 4.05 | 4.00 | 4.70 | 5.00 | 4.90 | 4.90 | 4.35 | 4.45 |
| 1898 | 4.00 | 4.27 | 4.17 | 4.15 | 4.80 | 4.50 | 4.17 | 4.20 | 4.15 | 4.00 | 3.85 | 3.75 |
| 1897 | 3.60 | 3.75 | 4.25 | 4.25 | 4.05 | 3.65 | 4.00 | 4.55 | 4.65 | 4.40 | 3.80 | 3.60 |
| 1896. | 4.45 | 4.35 | 4.35 | 4.15 | 3.75 | 3.60 | 3.70 | 3.75 | 3.50 | 3.65 | 3.67 | 3.65 |
| 1895 | 4.80 | 4.65 | 5.30 | 5.42 | 4.97 | 5.10 | 5.70 | 5.40 | 4.65 | 4.50 | 3.85 | 3.75 |

## Chicago Monthly Top Beef Cattle Prices.

| Years. | Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct.* | Nov. | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1916 | $9.85 | $ 9.75 | $10.05 | $10.00 | $10.90 | $11.50 | $11.30 | $11.50 | $11.50 | $11.65 | $12.40 | $13.00 |
| 1915 | 9.70 | 9.50 | 9.15 | 8.90 | 9.65 | 9.95 | 10.40 | 10.50 | 10.50 | 10.60 | 10.55 | 11.60 |
| 1914 | 9.50 | 9.75 | 9.75 | 9.55 | 9.60 | 9.45 | 10.00 | 10.90 | 11.05 | 11.00 | 11.00 | 11.40 |
| 1913 | 9.50 | 9.25 | 9.30 | 9.25 | 9.10 | 9.20 | 9.20 | 9.25 | 9.50 | 9.75 | 9.85 | 10.25 |
| 1912 | 8.75 | 9.00 | 8.85 | 9.00 | 9.40 | 9.60 | 9.85 | 10.65 | 11.00 | 11.05 | 11.00 | 11.25 |
| 1911 | 7.10 | 7.05 | 7.35 | 7.10 | 6.50 | 6.75 | 7.35 | 8.20 | 8.25 | 9.00 | 9.25 | 9.35 |
| 1910 | 8.40 | 8.10 | 8.85 | 8.65 | 8.75 | 8.85 | 8.60 | 8.50 | 8.50 | 8.00 | 7.75 | 7.55 |
| 1909 | 7.50 | 7.15 | 7.40 | 7.15 | 7.30 | 7.50 | 7.65 | 8.00 | 8.50 | 9.10 | 9.25 | 9.50 |
| 1908 | 6.40 | 6.25 | 7.50 | 7.40 | 7.40 | 8.40 | 8.25 | 7.90 | 7.85 | 7.65 | 8.00 | 8.00 |
| 1907 | 7.30 | 7.25 | 6.90 | 6.75 | 6.50 | 7.10 | 7.50 | 7.60 | 7.35 | 7.45 | 7.25 | 6.35 |
| 1906 | 6.50 | 6.40 | 6.35 | 6.35 | 6.20 | 6.10 | 6.50 | 6.85 | 6.95 | 7.30 | 7.40 | 7.90 |
| 1905 | 6.35 | 6.45 | 6.35 | 7.00 | 6.85 | 6.35 | 6.25 | 6.50 | 6.50 | 6.40 | 6.75 | 7.00 |
| 1904 | 5.90 | 6.00 | 5.80 | 5.80 | 5.90 | 6.70 | 6.65 | 6.40 | 6.55 | 7.00 | 7.30 | 7.65 |
| 1903 | 6.85 | 6.15 | 5.75 | 5.80 | 5.65 | 5.15 | 5.65 | 6.10 | 6.15 | 6.00 | 5.85 | 6.00 |
| 1902 | 7.75 | 7.35 | 7.40 | 7.50 | 7.70 | 8.50 | 8.85 | 9.00 | 8.85 | 8.75 | 7.40 | 7.75 |
| 1901 | 6.15 | 6.00 | 6.25 | 6.00 | 6.10 | 6.55 | 6.40 | 6.40 | 6.60 | 6.90 | 7.25 | 8.00 |
| 1900 | 6.60 | 6.10 | 6.05 | 6.00 | 5.85 | 5.90 | 5.85 | 6.20 | 6.15 | 6.00 | 6.00 | 7.50 |
| 1899 | 6.30 | 6.25 | 5.90 | 5.85 | 5.75 | 5.75 | 6.00 | 6.65 | 6.90 | 7.00 | 7.15 | 8.25 |
| 1898 | 5.50 | 5.85 | 5.80 | 5.50 | 5.50 | 5.35 | 5.65 | 5.75 | 5.85 | 5.90 | 6.25 | 6.25 |
| 1897 | 5.50 | 5.40 | 5.65 | 5.50 | 5.45 | 5.30 | 5.25 | 5.50 | 6.00 | 5.40 | 6.00 | 5.65 |
| 1896 | 4.00 | 4.75 | 4.75 | 4.75 | 4.55 | 4.65 | 4.60 | 5.00 | 5.30 | 5.30 | 5.45 | 6.50 |
| 1895 | 5.80 | 5.80 | 6.60 | 6.40 | 5.25 | 6.00 | 6.00 | 6.00 | 6.00 | 5.60 | 5.00 | 5.50 |

13. From the data of Exercise 12, construct the correlation table of hog prices and months of the year.

14. From data obtained from a financial journal construct a correlation table of the prices of common and preferred stocks.

15. In the correlation table of Exercise 12 does there appear to be a sharp tendency for the beef cattle arrays to vary with the changing live hog prices? Is the tendency more pronounced at some parts of the table than at others?

16. Compare the tendencies for close connection between the attributes in the table of Exercise 13 with that in Exercise 12.

**Correlation.** In the table of student heights and weights there is a decided tendency for heaviness and tallness to be associated and for lightness and shortness to be associated. There is likewise a pronounced tendency for the prices of live hogs and beef cattle to vary together. It is to be noted that the two series of measurements do not vary together in every case; that is, there are months in which the price of hogs is low but the price of beef high. But when all the months of an array are taken together the general tendency for the progressive increase of beef cattle prices with each increase of hog prices is evident. *Two characteristics are said to be* **correlated** *when there is a tendency for the changes in the value of one to depend on the changes in the value of the other.* The two characteristics may increase together or one may increase while the other decreases and even in a part of the table the movement of the changes may be together and in another part the two series of changes may move in opposition; the essential evidence for the presence of correlation is that the measurements change from array to array.

In **uncorrelated data** there is no tendency for the distributions of the arrays to change from type to type.

In **perfectly correlated** data there is an *exact* connection between the values of the two characteristics. If height and weight were perfectly correlated, for instance, all persons of a given height, say 68 inches, would be of the same weight and hence all the frequencies of the weight array of type 68 would lie within a single subgroup. Between the two extremes of *perfect* and of *no* correlation there are all degrees of correlation.

### Exercises.

17. Study the degrees of correlation shown by the tables constructed in working the exercises of this chapter.

18. Is it possible to find actual data which shows absolutely no correlation? Construct an imaginary table which shows no correlation.

## CHAPTER VIII.

### THE CORRELATION RATIO.

**The Mean as Representative of the Array.** In Chapter IV it was stated that the modal deviation is the most frequent deviation; that is, the most typical deviation of a distribution. Because the mode cannot be computed by a simple and uniform process of arithmetic the mean is a more practicable representative of the array. And this substitution of the mean for the mode will rarely produce a serious error.

Since the mean of the frequencies of an array is taken as the representative of the deviations of the array, from the definition of correlation on page 73 it is apparent that the amount or degree of correlation in the data will be indicated by the variation in the means from array to array.

**Regression Curves.** The variation in the means of the arrays is shown graphically by the curve of means, which is called **a regression* curve.**

Since there are two sets of arrays there are two regression curves.

**Coördinate Axes.** It is usual to take for the horizontal or $x$-axis the horizontal line thru the mean of all the $y$'s; that is, the horizontal line at a distance $\bar{y}$ below the base line of the table, and for the $y$-axis the vertical line distant $\bar{x}$ from the left marginal vertical. The point of intersection of these two lines is called the **center** of the table. Deviations to the right are taken positive and those to the left negative; deviations downward from the new horizontal axis are positive and deviations upward are considered negative. Sometimes this convention of plus downward and negative upward is departed from. No confusion can result however if it is remembered that the directions in which an attribute is increasing is always taken as positive.

* So called by Francis Galton for certain reasons which arose in his investigations in biology. The name has become general.

### Exercises.

1. Draw the axes and regression curves for each of the correlation tables of Chapter VII.

2. Study and compare the forms of the regression curves of Exercise 1.

**Correlation and the Regression Curves.** In uncorrelated data the means of an array does not depend on the type of the array; that is, does not change from array to array, and hence the unchanging value of the means must be the same as the mean of all the $y$'s.

The **regression curve for uncorrelated data** therefore approximates a straight line coinciding with the horizontal axis. For correlated data the regression curve diverges or deviates from this position of coincidence with the axis. It must be noted that the shape of the regression curve may be quite irregular without effect on the degree of correlation present in the data; it is the distance of the means from the axis that counts in determining the degree of correlation present. Hence any numerical measure of the extent of correlation in the data must depend on the deviation of the means from the horizontal axis thru the center.

Since there are two regression curves and two axes there are two correlations in each correlation table and their numerical measures involve the deviations of the respective regression curves from the corresponding straight lines thru the center. Thus the dependence of height on weight and of weight on height are two distinct correlations.

**Mean Squared Deviation of the Means of Arrays.** The mean squared deviation is the most convenient measure of the deviation of the means of the arrays. In computing this the means of the arrays are first written in a vertical column and then the difference between each mean and the mean of all the variates is set down in a second column. Because the differences are used only in the squared form it is not necessary to retain a negative sign.

The third column in the computations of Table X, page 77, contains the squares of the differences. Since the means of the array are used as the representatives of the individuals of the respective arrays each of these individuals is possessed of

the squared deviations. Hence each square must be multiplied by the respective frequency of the corresponding array. The resultant products form the fourth column. The sum of this fourth column is the total sum of squared deviations and this sum divided by the total frequency is the **mean squared deviation.**

**The Correlation Ratio.** The mean squared deviation just obtained would be a significant measure of correlation were it not for the fact that it does not take into account the dispersion of the data as a whole Without changing the mean and the frequency of a single $y$-array, it would be possible to spread out each array to twice its length. This alteration would concern the dispersion of the data as a whole but would leave the mean square deviation from the horizontal axis unchanged. It is evident that the value of the mean square deviations of the means of the arrays is of less significance in the more spread out data. Hence the dispersion of the data as a whole must be considered in interpreting the value of the mean squared deviation. The dispersion of the data as whole is given by the standard deviation of the frequencies of the totals in the vertical sum column. The smaller this mean square deviation the more significant is the deviation of the means, and the larger this standard deviation the less significant the deviation of the means. It is therefore reasonable to divide the square root of the mean square deviation of the means of the arrays by the standard deviation from the marginal column. The quotient is called the **correlation ratio,** and is denoted by the Greek letter $\eta$.

The computation of the correlation ratio for the dependence of student weight on height follows.

A carefully planned outline scheme of computation must be made before the figures are entered.

The means and the one standard deviation were computed in the usual manner. We have, for the data as a whole, $\bar{y} = 7.9$, $\sigma^2 = 9.79$. The means of the arrays are written in the second column just after the frequencies. The differences between the means and $\bar{y}$ follow in the third column. The squares, and the product of the squares by the frequencies are the fourth and fifth columns respectively. The symbols explained in Chapter VII are written at the head of each column.

**Computation of $\eta$.**

| $n_x$ | $\bar{y}_x$ | $\bar{y}-\bar{y}_x$ | $(\bar{y}-\bar{y}_x)^2$ | $n_x(\bar{y}-\bar{y}_x)^2$ |
|---|---|---|---|---|
| 2 | 9.5 | 1.6 | 2.56 | 5.12 |
| 10 | 4.7 | 3.2 | 10.24 | 102.40 |
| 11 | 4.4 | 3.5 | 12.25 | 134.75 |
| 38 | 4.6 | 3.1 | 9.61 | 365.18 |
| 57 | 6.1 | 1.8 | 3.24 | 184.68 |
| 93 | 6.8 | 1.1 | 1.21 | 112.53 |
| 106 | 6.9 | 1.0 | 1.00 | 106.00 |
| 126 | 8.0 | 0.1 | 0.01 | 1.26 |
| 109 | 8.8 | 0.9 | 0.81 | 89.19 |
| 87 | 9.7 | 1.8 | 3.24 | 281.88 |
| 75 | 10.9 | 3.0 | 9.00 | 675.00 |
| 23 | 10.3 | 2.4 | 5.76 | 132.48 |
| 9 | 11.1 | 3.2 | 10.24 | 92.16 |
| 4 | 10.5 | 2.6 | 6.76 | 27.04 |
| 750 | | | | 2309.67 |

$$\eta^2 = \frac{2309.67}{750 \times 9.79} = 0.3146$$

$$\eta = 0.56$$

TABLE X.

## Exercises.

3. Compute the value of $\eta$ for the dependence of monthly precipitation upon monthly mean temperature as shown by the data of the Columbus Weather Station.

4. Compute the value of $\eta$ for the correlation of Chicago top hog prices with Chicago top beef cattle prices as shown in the table of Exercise 12 of the preceding chapter.

5. Compute values of $\eta$ from the tables of Exercises 13 and 14 of Chapter VII.

**Two Values for $\eta$ in Each Table.** From the method of computation it is clear that there are two values for $\eta$ in each correlation table, one for each regression curve. The correlation ratio of weight with height, for instance, may differ considerably from the correlation ratio of height with weight; the dependence of precipitation on temperance may be of a decidedly different degree from that of temperature on precipitation. The two values of $\eta$ do not ordinarily differ markedly but there can be no apriori assurance that they will be essentially of equal value and hence it is necessary to compute the two values separately in case both are desired. To distinguish the two

measures, for the dependence of $y$ on $x$, of weight on height, the symbol $\eta_y$ is used and the symbol $\eta_x$ refers to the dependence of $x$ on $y$.

### Exercises.

6. Compute the value of $\eta$ for the correlation of height with weight and compare with the other value of $\eta$ computed on page 77.

7. Compute the value of $\eta_x$ from the precipitation-temperature correlation table, and compare the values of $\eta_x$ and $\eta_y$.

8. Compute the value of $\eta_x$ for the live stock price table of Exercise 12, Chapter VII, and compare with the value of $\eta_y$ from the same table.

**Limiting Values of the Correlation Ratio.** In theory, the means lie exactly on the axis for data of zero correlation. Each separate item, therefore, in the mean square deviation of the means is zero and hence $\eta$ **is zero for zero correlation.**

Because each term of the mean squared deviation of the means is squared and hence necessarily positive any accidental fluctuations of the means of the arrays in data of essentially zero correlation increase the value of $\eta$. Since there are no compensating flunctuations, the result is that small values of $\eta$ are quite likely to be too large and hence the statistical significance of $\eta$ for data of a small degree of correlation is open to question. The degree of correlation in such cases cannot be greater than the value of $\eta$ would indicate and it may be less. It must be evident from the nature of the error that for material showing a considerable degree of correlation the error from this source is negligible.

As discussed in Chapter VII, in perfectly correlated data the frequencies of each array are concentrated in a single class or subgroup of the array. According to the shape of the regression curve two cases can arise for perfectly correlated data, as the following imaginary distributions illustrate.

| | | | |
|---|---|---|---|
| 1 | | | |
| | 3 | | |
| | | 6 | |
| | | | 4 |

| | | | |
|---|---|---|---|
| 1 | 3 | | |
| | | 0 | |
| | | | 4 |
| | | 6 | |

In the first distribution it is evident that the mean squared deviation of the means is obtained by exactly the same numerical work and by the use of the same numbers as is the standard deviation of all the $y's$ and hence $\eta$, which is the ratio of these two measures, must be equal to unity. A study of the second distribution leads to the conclusion that in this case also the two measures of deviation are equal and hence $\eta$ is unity as it is in the first distribution.

### Exercises.

9. Compare the values of $\eta$ that have been computed with the general appearance of correlation in the tables.

10. Can a tendency be detected for the two values of $\eta$ to be closer together in value for highly correlated data than for data of smaller correlation?

11. Describe in general terms the two species of perfect correlation illustrated above.

**Probable Deviation of the Correlation Ratio.** It can be proved that the probable deviation of a correlation ratio is given by the formula

$$P.\ E.\ \eta = 0.6745\ \frac{(1-\eta^2)}{\sqrt{N}}$$

### Exercises.

12. Compute the probable deviations of the correlation ratios of this chapter.

The probable error is of much practical use in estimating the significance of values of $\eta$ tho of course the facts of its theoretical derivation must always be kept in mind. It is assumed in the derivation of the formula that the data is strictly homogeneous thruout; that all the fluctuations from sample to sample are merely those of random sampling; that the regressions are truly linear; and that each array has the same spread.

In working with correlations, especially where the total frequencies are not large, it is always well to obtain a considerable number of distributions. Then if there proves to be a consistency in the value of $\eta$ greater confidence can be placed in those values than if there was only one distribution. Thus if fifty groups of 750 students were each measured for height

and weight and the computed values for $\eta$ should show a decided tendency to agree in value, increased weight could be given to these values of $\eta$.

**Spurious Correlation.** In interpreting the computed value of any measure of correlation care must be taken that the correlation is not merely apparent and due to the method of obtaining the data. Wages and general prices appear to be highly correlated but how much of this apparent connection is due to the fact that both are expressed in terms of money which has cheapened and consequently, at least to some extent, caused both wages and general prices to increase. When money is becoming cheaper both wages and prices tend in general to rise together; when money is becoming dearer both wages and prices tend in general to fall together. Such variations in the purchasing value of money could thus introduce a considerable element of apparent connection between the two attributes—wages and general prices — even when there was no tendency whatever for real wages to change, that is, no change in the amount of goods that the laborer could purchase with his wages.

## Exercises.

13. In which of the correlation of this chapter is there a possibility of spurious correlation?

14. Show that in correlating index numbers especial care is necessary in interpreting values of $\eta$.

15. Show that where there is an element of spurious correlation present the correlation is real in so far as the measurements themselves are concerned.

## CHAPTER IX.

### THE COEFFICIENT OF CORRELATION.

**Linear Regression.** A straight line *fitted* to the means of the arrays is called a **line of regression.** A line of regression smooths the curve of regression. Whenever a curve of means approximates a straight line the regression is said to be **sensibly linear.** If the regression curve, within the limits of accuracy of the data, is *exactly* a straight line the regression is said to be **truly linear.**

The slope of a regression line shows the broad general tendencies of the connection between the attributes. Does weight tend to increase as height increases? Does the monthly precipitation increase with an increase of temperature? If so at about what rate? These are questions that depend for an answer on the slopes of the regression curves. It may happen that in some correlation tables the regression curves deviate in form so widely from a straight line that the line of regression has but little significance; in such cases the usefulness of any statement of general tendencies is open to question.

#### Exercises.

1. Draw by inspection the regression lines on the correlation table of student heights and weights.

2. In Exercise 1 estimate the comparative degrees of correlation shown by the two regression lines.

**The Equations of the Lines of Regression.** Let the coordinate axes be the two lines thru the center determined by the means of all the variates as described on page 74. Then $\bar{y}_x$ and $x$ are the coordinates of a point on the one regression line and $\bar{x}_y$ and $y$ on the other. It must be understood, however, that the values of $\bar{y}_x$ and $\bar{x}_y$ here referred to are the adjusted or fitted means of the arrays so that unless the regressions are truly linear these values will differ from the values obtained by actual computation.

It is demonstrated in Chapter XII that the equation of the regression line of the means of the $y$ arrays is

$$\bar{y}_x = r\frac{\sigma_y}{\sigma_x}x.$$

And of the means of the $x$ arrays,

$$x_y = r\frac{\sigma_x}{\sigma_y}y;$$

where $\sigma_y$ is the right hand marginal standard deviation and $\sigma_x$ the bottom marginal standard deviation. The constant $r$ is defined by the equation, $r = \dfrac{\Sigma\Sigma n_{xy}.\, x.y}{N\sigma_x \,.\, \sigma_y}$. The expression $\Sigma\Sigma n_{xy}.xy$ is a symbolic way of saying: the sum obtained by multiplying the frequency of each class by its deviation from the horizontal axis and then by its deviation from the vertical axis and then adding the products.

According to the first of the two regression equations the mean weight for height 71 is obtained by substituting the value for x measured from the mean and multiplying and dividing as the formula directs. We found that for this data of student measurements $\sigma_y = 2.36$ and $\sigma_x = 3.14$. The value of $r$ is found presently to be 0.56. The array is distant from the mean 3.1. Hence,

$$\bar{y}_{8:1} \text{ or } \bar{y}_{71} = 0.56\frac{2.36}{3.14}3.1$$
$$= 1.31, \text{ weight classes from the mean weight.}$$

This use of the regression lines to estimate the position of the means is often of practical value.

**The Coefficient of Correlation.** Let us now compute the correlation ratio using, however, in case the regression is not truly linear, *not the actual means of the array but the means given by the regression line.* The deviation of a mean from the horizontal axis is $r \,.\, \dfrac{\sigma_y}{\sigma_x} \,.x$. The square of this quantity multiplied by the frequency of the array is $n_x x^2 \left(\dfrac{r^2\sigma_y^2}{\sigma_x^2}\right)$

The last factor is the same for each array and the sum of the other factors leads to the standard deviation of all the $x$'s. Hence we have, on carrying out the multiplications for each array and adding, $r^2 \frac{\sigma_y^2}{\sigma_y^2} \Sigma n_x \cdot x^2 = \frac{r^2\sigma_y^2}{\sigma_x^2} \cdot N\sigma_x^2 = Nr^2\sigma_y^2$. Therefore the mean squared deviation of the regression means of the arrays is $\frac{Nr^2\sigma_y^2}{N} = r^2\sigma_y^2$. On dividing this mean squared deviation by the standard deviation of all the $y$'s we have $\frac{r^2\sigma_y^2}{\sigma_y^2} = r^2$. That is, the constant $r$ turns out to be the correlation ratio when the regression means are used instead of the true means. It is called the **Coefficient of Correlation.**

**Computation of r.** For computation purposes the summation $\Sigma_x\Sigma_y n_{xy} yx$ can be arranged in the following manner. Let the subgroup frequencies of a given $y$ array be each multiplied by the respective deviations, all deviations being measured from the axis thru the center, and the products summed. Divided by the frequency of the array this sum gives the mean $\bar{y}_x$. Hence the summation for the array is equal to the product of the mean $\bar{y}_x$ and the frequency $n_x$. On making this substitution the original summation formula becomes $\Sigma n_x . \bar{y}_x . x$, or $\Sigma n_x(\bar{y}_x - \bar{y})\ (x - \bar{x})$ from the original axes.

In the course of the computation of the correlation ratio the means $\bar{y}_x$ are obtained and hence to the computation schedule of page 77 only the additional column for the $x$ deviations of each array is needed. Then the multiplication of the corresponding values from the $n_x$, $(\bar{y}_x - \bar{y})$, and $(x - \bar{x})$ columns gives the column which sums into the quantity $\Sigma n_x(\bar{y}_x - \bar{y})(x - \bar{x})$. This sum divided by the product of the three factors $N$, $\sigma_x$, and $\sigma_y$ gives the required value for $r$.

Table XI which follows shows the computation. The value of $\sigma_y$ is found in Exercise 10, Chapter V, to be 2.31, and $\sigma_x = 3.13$.

**Computation of r.**

| $x$ | $\bar{y}_x$ | $\bar{y}_x-\bar{y}$ | $n_x$ | $x-\bar{x}$ | $n_x(x-\bar{x})$ | $n_x(x-\bar{x})(\bar{y}_x-\bar{y})$ |
|---|---|---|---|---|---|---|
| 1 | 9.5 | —1.6 | 2 | —6.9 | —13.8 | 22.08 |
| 2 | 4.7 | —3.2 | 10 | —3.9 | —29.0 | 92.8 |
| 3 | 4.4 | —3.5 | 11 | —4.9 | —53.9 | 188.65 |
| 4 | 4.6 | —3.1 | 38 | —3.9 | —148.2 | 459.42 |
| 5 | 6.1 | —1.8 | 57 | —2.9 | —165.3 | 297.54 |
| 6 | 6.8 | —1.1 | 93 | —1.9 | —176.7 | 194.37 |
| 7 | 6.9 | —1.0 | 106 | —0.9 | —93.4 | 93.4 |
| 8 | 8.0 | +0.1 | 126 | +0.1 | +12.6 | 1.26 |
| 9 | 8.8 | +0.0 | 109 | +1.1 | +119.9 | 0. |
| 10 | 9.7 | +1.8 | 87 | +2.1 | +182.7 | 328.86 |
| 11 | 10.9 | +3.0 | 75 | +3.1 | +232.5 | 697.5 |
| 12 | 10.3 | +2.4 | 23 | +4.1 | +94.3 | 226.32 |
| 13 | 11.1 | +3.2 | 9 | +5.1 | +45.9 | 146.88 |
| 14 | 10.5 | +2.6 | 4 | +6.1 | +24.4 | 63.44 |
| | | | | | $\Sigma n_x(x-\bar{x})\ (\bar{y}-\bar{y})$ | $=3018.4$ |

$$r=\frac{\Sigma n_x\ (\bar{y}_x-\bar{y})\ (x-\bar{x})}{N\sigma_x\ \sigma_y}$$

$$=0.55$$

TABLE XI.

## Exercises.

3. Compute the value of $r$ for the monthly precipitation and temperature data.

4. Compute $r$ for the top-hog and top-beef-cattle data.

5. Compare the values of $r$ in Exercises 3 and 4 and in the weight-height data with the corresponding values for $\eta$.

6. Compute the value of $r$ from the monthly price-of-hogs data of Exercise 12, Chapter VII. Compare with the corresponding value for $\eta$.

7. Does there seem to be a tendency for $\eta$ and $r$ to agree more closely for highly correlated data than for material of small correlation?

8. Compare the amount of labor involved in the computation of $r$ with that involved in the computation of $\eta$.

**The Relation of r to $\eta$.** In data exhibiting a regression which is truly linear the value of $\eta$ is, of course identical with that of $r$. In the case of any but truly linear regression it is readily shown in Chapter XII that the value of $r$ is necessarily less than that of $\eta$. In fact if the regression curve is of a

certain shape the value of $r$ will be very small even tho practically perfect correlation exists.

Unlike the correlation ratio the coefficient of correlation expresses a property of the correlation table as a whole and not merely of one or the other of the two correlations of the table.

Again, unlike the correlation ratio, the negative sign obtained in the final extraction of the square root (in the discussion of page 76) has a significance; it indicates that the regression line has a negative slope and hence that the connection between the attributes is **inverse**; that is, one attribute increases while the other decreases.

Because both positive and negative values of $r$ can occur there is no tendency, as there is in the case of $\eta$, for small values of $r$ to be larger than the actual degree of correlation would warrant.

**Limiting Values for** $r$. In data of zero correlation it is clear that the regression line coincides with the axis and hence the value of $r$ must be zero.

Reasoning from the relation of $r$ to $\eta$ we see that for truly linear regression perfect correlation leads to a value of $r$ equal to unity. The unity value for $r$ will be positive or negative according to the correlation is direct or inverse. According to the underlying theory of the coefficient of correlation for data in which a regression is not linear the value of $r$ cannot be unity even tho there is perfect correlation and hence $r$ is necessarily smaller in value than the degree of correlation would require.

**Statistical Properties of the Coefficient of Correlation.** The coefficient $r$ is, as the preceding discussions show, a conservative measure of correlation. In periodic data exhibiting a sinusoid form for the regression curve the correlation may be high but because the departure of the regression from linearity is so wide the value of $r$ understates the correlation and hence its applicability in such data is not of importance.

The characteristic importance of the coefficient $r$ is in defining the slope of the regression lines. It furnishes the most convenient method for defining the general tendencies in the data. The rise of prices, for instance, during the last fifteen years can be readily measured by the rate of rise of the regression line.

Therefore for the single purpose of measuring correlation the coefficient of correlation is distinctly inferior to the correlation ratio both in convenience and reliability. It should never be used as a measure of correlation without first carefully testing the form of the regression. It does have however the highly useful property of giving the slopes of the regression lines.

**Test for Linearity of Regression.** It would be suspected from the preceding theory and discussion that the difference between $\eta$ and $r$ should be an indicator of the departure of the regression from linearity. A somewhat more convenient measure of this departure than the simple difference is the difference of the squares of $\eta$ and $r$.

**Probable Deviations.** The following probable deviations can be derived.

$$P.\ E.\ \text{of}\ r = 0.6745\ \frac{(1-r^2)}{\sqrt{N}}$$

$$P.\ E.\ \text{of}\ (\eta^2 - r^2) = \frac{1.35}{\sqrt{\eta}}\sqrt{\eta^2 - r^2}$$

A practical criterion of linearity is to assume linearity when

$$\frac{\sqrt{N}}{1.35}\sqrt{\eta^2 - r^2} < 2.5.$$

8. Compute the regression equations for each of the correlation tables of Chapter VII.

10. How can the value of $r$ be obtained graphically from the regression lines? Is this a practicable method of finding the value of $r$?

11. Compute the measure of departure from linearity, $(\eta^2 - r^2)$ for the correlation tables of Chapter VII.

12. A correlation table has two measures of departure from linearity. Show that one regression may be linear and the other nonlinear.

13. Show that if the value of $r$ is high the regressions must both be approximately linear.

14. By extending the regression line estimate the price of live hogs for January, 1917.

15. What weight should correspond to a student height of 50 inches?

16. What is the best estimate of the temperature for a month with a precipitation of 3.4 inches?

17. Discuss the value of the probable deviations in exercises 14-16.

## CHAPTER X.

### CORRELATION FROM RANKS.

**Rank in a Series.** When the data consists not of the direct measurements of the characteristics but of their order or rank in a series the correlation of the ranks may differ materially from the true variate correlation. Let us define **rank** as position in a series so that an individual of rank *one* would have no individuals above or before it; an individual of rank *two* would have one individual before it, etc.

To pass from rank to variate correlation it is necessary to know the form of the distribution of the values of the characteristics. Only for normal distributions has the requisite theory been developed. It is consequently necessary to employ the same formulas for other forms of distributions, although this may sometimes open the way to serious inaccuracies.

Let the ranks of the same individual in regard to the respective characteristic be $\nu_x$ and $\nu_y$. Let there be $N$ individuals and let $\gamma_x$ and $\gamma_y$ denote the respective means of the two series and $\sigma_{\nu_x}$ and $\sigma_{\nu_y}$ the standard deviations.

Also let all the measurements of each characteristic be distinct in value; that is, let there be no equal measurements.

**Theorem I.** *The mean ranks* $\gamma_x$ *and* $\gamma_y$ *are each equal to* $(N+1)/2$.

Since there are as many ranks as individual measurements and since the ranks proceed uniformly from 1 to $N$ the mean is $(N+1)/2$.

**Theorem II.** *The standard deviations of the ranks are each equal to* $\frac{N}{12}$.

For, $N\sigma_{\nu_x}^2 = \Sigma(\gamma_x - \nu_x)^2$,

$$= \Sigma\gamma_x^2 - 2\nu_x\Sigma\gamma_x + \Sigma\nu_x^2,$$

$$= \frac{1}{6}N(N+1)(2N+1) - 2\gamma_x \cdot \frac{N(N+1)}{2} + N\gamma_x^2,$$

on applying the rules $\Sigma N^2 = 1/6N(N+1)(2N+1)$ and $\Sigma N = 1/2N(N+1)$,

$$= 1/6N(N+1)(2N+1) - \frac{N(N+1)^2}{2} + \frac{N(N+1)^2}{4},$$

$$= \frac{1}{12}(N^3 - N).$$

Therefore $\sigma_{v_x}^2 = \frac{1}{12}(N^2 - 1)$.

The following theorem is necessary for the computation of rank correlation.

**Theorem III.** *If* $\sigma_x = \sigma_y$, $r = 1 - \frac{\sigma^2_{(x-y)}}{\sigma_x^2}$.

For, $(x-y)^2 = x^2 + y^2 - 2xy$,
or $N\sigma^2_{(x-y)} = N\sigma_x^2 + N\sigma_y^2 - 2\Sigma_{xy}$.
and $\Sigma(x-y)^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$,
But $\Sigma xy = r \cdot N \cdot \sigma_x \sigma_y$.
Therefore $N\sigma^2_{(x-y)} = N\sigma^2 - 2Nr\sigma_x\sigma_y$,

and $$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma^2_{(x-y)}}{2\sigma_x\sigma_y}.$$

If $\sigma_x = \sigma_y, r = 1 - \frac{\sigma^2_{(x-y)}}{2\sigma_x^2}$.

## Exercises.

1. Show how to compute the value of $r$ from the data of Table VIII by the formula $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma^2_{(x-y)}}{2\sigma_x\sigma_y}$.

**Theorem IV.** *The correlation coefficient of the ranks* $v_x$ *and* $v_y$ *is given by the formula,*

$$r_{v_x v_y} = 1 - \frac{6\Sigma(v_x - v_y)^2}{N(N^2-1)}.$$

On making use of theorem 3, we have,

$$r_{v_x v_y} = 1 - \frac{\sigma^2_{(x-y)}}{2\sigma_{v_x}^2},$$

$$= 1 - \frac{\Sigma(v_x - v_y)^2}{2N\sigma_{v_x}^2}$$

$$= 1 - \frac{\Sigma(v_x - v_y)^2}{2 \cdot \frac{1}{12} \cdot N(N^2-1)}, \qquad \text{from theorem 2,}$$

$$= 1 - \frac{\Sigma(v_x - v_y)^2}{\frac{1}{6}N(N^2-1)}.$$

To illustrate the method let us compute the rank correlation between yearly mean temperature and yearly mean rainfall for Ohio as shown by the data of Exercises 2 and 3 of Chapter II. The yearly means are obtained by adding the monthly means and dividing by twelve.

The order of the twenty-four years in respect to temperature is written in the first column and in respect to rainfall in the second. The ties are disposed of by assigning the ranks in the inverse order of the time, thus 1903 and 1902 at 50.5 and 1903 is given rank 15 and 1902, 16. But the matter of ties will be presently considered. The third column contains for each year the differences in rank with respect to the two attributes, temperature and rainfall, and the fourth the squared differences. On adding the fourth column and applying the formula $r = 1 - \frac{6\Sigma(v_x - v_y)^2}{N(N^2-1)}$ we find $r = 0.10$.

| *Year.* | *Temp.* | *Rainfall.* | *Difference.* | *Sq. Diff.* |
|---|---|---|---|---|
| 1911 | 1 | 4 | 3 | 9 |
| 1910 | 17 | 6 | 11 | 121 |
| 1909 | 13 | 5 | 8 | 64 |
| 1908 | 5 | 19 | 14 | 196 |
| 1907 | 22 | 3 | 19 | 361 |
| 1906 | 9 | 14 | 5 | 25 |
| 1905 | 20 | 10 | 10 | 100 |
| 1904 | 24 | 17 | 7 | 49 |
| 1903 | 15 | 16 | 1 | 1 |
| 1902 | 16 | 13 | 3 | 9 |
| 1901 | 18 | 24 | 6 | 36 |
| 1900 | 2 | 21 | 19 | 361 |
| 1899 | 11 | 18 | 7 | 49 |
| 1898 | 6 | 2 | 4 | 16 |
| 1897 | 14 | 11 | 3 | 9 |
| 1896 | 7 | 9 | 2 | 4 |
| 1895 | 21 | 23 | 2 | 4 |
| 1894 | 3 | 22 | 19 | 361 |
| 1893 | 10 | 7 | 3 | 9 |
| 1892 | 19 | 14 | 5 | 25 |
| 1891 | 8 | 12 | 4 | 16 |
| 1890 | 4 | 1 | 3 | 9 |
| 1889 | 11 | 20 | 9 | 81 |
| 1888 | 23 | 8 | 15 | 225 |

$N = 24$ $\qquad \Sigma(\nu_x - \nu_y)^2 = 2{,}150$

$N(N^2-1) = 13{,}800 \qquad 6\Sigma(\nu_x - \nu_y)^2 = 12{,}900$

$$r = 1 - \frac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2-1)},$$

$$= 0.10.$$

**Ties in Rank.** The application of the formula $r_{\nu_x \nu_y} = 1 - \dfrac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2-1)}$ is straightforward and direct. The only uncertainty arises from ties in the measurements. Thus in the preceding illustrative example it is found that the temperature for each of the two years 1907 and 1894 is 52.3. What ranks are to be assigned to each of the measurements? In order to avoid complicating details in an illustrative problem, in the preceding computation we gave the later year the numerically smaller rank, but ordinarily it is better to base the ranks on one of the two plans:

(1). **The Bracket Rank method,** under which the ties are assigned the same rank and that equal rank is taken as the rank next greater than that of the individual preceding the ties. The next individual after the ties takes the same rank as if preceding ties had each been given ranks differing by unity. Thus under this method the ranks of the illustrative example are as given in the table below.

(2). **The Mid-Rank method,** under which all ties are given the same rank but that rank is the rank of the mid-individual. In the column below the two methods may be compared.

Under either method the total number of ranks must be the same and equal to $N$.

| | *Temperature.* | *Bracket Method.* | *Mid Rank Method.* |
|---|---|---|---|
| 1911 | 52.6 | 1 | 1 |
| 1900 | 52.3 | 2 | 3 |
| 1894 | 52.3 | 2 | 3 |
| 1890 | 52.3 | 2 | 3 |
| 1908 | 52.1 | 5 | 5.5 |
| 1898 | 52.1 | 5 | 5.5 |
| 1892 | 51.7 | 7 | 7.5 |
| 1891 | 51.7 | 7 | 7.5 |
| 1906 | 51.6 | 9 | 9.5 |
| 1893 | 51.6 | 9 | 9.5 |
| 1899 | 51.5 | 11 | 11 |
| 1889 | 51.1 | 12 | 12 |
| 1909 | 50.9 | 13 | 13 |
| 1897 | 50.6 | 14 | 14 |
| 1903 | 50.5 | 15 | 15.5 |
| 1902 | 50.5 | 15 | 15.5 |
| 1910 | 50.4 | 17 | 17 |
| 1901 | 50.2 | 18 | 18 |
| 1892 | 50.1 | 19 | 19 |
| 1905 | 50.0 | 20 | 20 |
| 1895 | 49.9 | 21 | 21 |
| 1907 | 49.6 | 22 | 22 |
| 1888 | 49.5 | 23 | 23 |
| 1904 | 48.6 | 24 | 24 |

2. Compute $r_{\nu_x \nu_y}$ from the above "bracket method" ranks.
3. Compute $r_{\nu_x \nu_y}$ from the above "mid-rank method" ranks.

**Probable Deviation of the Rank Coefficient.** As given by Pearson;

$$P.\ E.\ \text{of}\ r_{v_x v_y} = \frac{0.6745}{\sqrt{N}}\ (1 - r_{v_x v_y}).$$

**Perfect Rank Correlation.** The ranks are perfectly correlated, according to the formula, when $\Sigma(v_x - v_y)^2 = 0$; that is, when each individual has the same rank in both series. Also there is perfect negative correlation when temperature and rainfall are inversely related so that the year with the highest temperature is the year with the lowest rainfall and so on up to the year with the lowest temperature which is associated with the highest rainfall. In this case of perfect negative correlation when $N$ is odd,

$$\Sigma(v_x - v_y)^2 = .\ 2\{(N-1)^2 + (N-3)^2 + \ldots + 2^2 + 0\},$$

$$= \left\{0 + 1 + 4 + 9 + \ldots \frac{(N-1)^2}{2}.\right\},$$

$$= 8 \cdot \tfrac{1}{6} \cdot \frac{N-1}{2} \cdot \frac{N+1}{2} \cdot N,$$

$$= \tfrac{1}{3} N(N^2 - 1),$$

Therefore $r = 1 - \dfrac{2N(N^2-1)}{N(N^2-1)}$,

$= -1$, a result according with the usual idea of inversely correlated attributes.

**Uncorrelated Data.** According to the formula, the sum of the squares of the differences of the ranks is equal to the sum of the squares of the ranks when $r = 0$. Thus when $r = 0$ subtracting the ranks has lost its significance — and this is exactly the idea of zero correlation.

Hence the *rank* coefficient $r$, is accurately significant for both perfect and zero correlation.

**A Correction Formula for the Rank Coefficient.** There is no assurance however that in general the *rank* $r$ will exactly express the true variate correlation. For instance, note the two following series of deviations,

100, 80, 70, 65, 62, 60, 55, 50, 40, 20; and 100, 99, 98, 97, 96, 95, 10, 9, 8, 7.

The ranks are the same in each series, namely,,

1, 2, 3, 4, 5, 6, 7, 8, 9, 10 .

The coefficient $r_{\nu_x \nu_y}$ which depends solely on the ranks, has the same value for a series of which the first is typical as it does for a series of which the second is typical. And yet the two distributions are fundamentally distinct in form.

Therefore, except for the two extreme cases of material of very high and of very low correlation, the value of a correlation constant computed from ranks must be interpreted with caution.

For a distribution which is approximately normal in form the following correction formula for $r_{\nu_x \nu_y}$ has been derived by Pearson.*

$$r_{xy} = 2 \sin \frac{\pi}{6} \cdot r_{\nu_x \nu_y}.$$

From Table X the values of $r_{xy}$ can be obtained directly from the value of $r_{\nu_x \nu_y}$ for each 0.05 of $r_{\nu_x \nu_y}$.

**Corresponding Values of $r_{xy}$ and $r_{\nu_x \nu_y}$.**

| $r_{\nu_x \nu_y}$ | $r_{xy}$ | $r_{\nu_x \nu_y}$ | $r_{xy}$ |
|---|---|---|---|
| 0.00 | 0.00 | 0.55 | 0.57 |
| 0.05 | 0.06 | 0.60 | 0.62 |
| 0.10 | 0.10 | 0.65 | 0.67 |
| 0.15 | 0.16 | 0.70 | 0.72 |
| 0.20 | 0.20 | 0.75 | 0.77 |
| 0.25 | 0.26 | 0.80 | 0.87 |
| 0.30 | 0.31 | 0.85 | 0.86 |
| 0.35 | 0.36 | 0.90 | 0.91 |
| 0.40 | 0.42 | 0.95 | 0.96 |
| 0.45 | 0.47 | 1.00 | 1.00 |
| 0.50 | 0.52 | .... | .... |

TABLE X.

For other values of $r_{\nu_x \nu_y}$ the corresponding values of $r_{xy}$ are readily obtained by interpolation.

*"On Further Methods of Determining Correlation" Drapers Co. Research Memoirs: Bio. Ser. IV.

**Probable Deviation of $r_{xy}$ Computed from Ranks.** As given by Pearson*

$$P.\ E.\ \text{of}\ r_{xy}\ \text{from ranks} = \frac{0.7063}{\sqrt{N}}(1 - r^2)\ .$$

### Exercises.

4. Determine $r_{xy}$ from the value of $r_{\nu_x\nu_y}$ of Exercise 1.

5. Compute the value of the rank $r$ from the data of other exercises and compare with the computed values of the variate $r$.

**Theorem V.** *Multiplying each rank $\nu_x$ or $\nu_y$ by a constant does not change the value of $r$.*

For if each rank is multiplied by $m$, the means $\nu_x$ and $\nu_y$ are each multiplied by $m$ so that the standard deviations are each multiplied by $m^2$. Also $\Sigma(\nu_x - \nu_y)^2$ is multiplied by $m^2$ and hence $r$ is unchanged.

**Theorem VI.** *Multiplying the ranks of one column but not of the other in general produces a change in the value of $r_{\nu_x\nu_y}$*

In this case the formula $r = 1 - \dfrac{6\Sigma(\nu_x - \nu_y)^2}{N(N^2 - )}$ cannot be used and the formula

$$r = \frac{N\sigma_{\nu_x}^2 + N\sigma_{\nu_x}^3 - \Sigma(\nu_z - \nu_z)(\nu_y - \nu_y)}{2r\sigma\nu_x\sigma\nu_y}$$

must be employed.

**The Accuracy of the Coefficient $r_{xy}$ when computed from Ranks.** *When the measurements are arranged in ranks and the coefficient is computed from the ranks alone, the computation is based on the relatively limited information which the ranks can convey. Hence the resulting coefficient can not be as trustworthy and reliable as the moment cofficient. However, when a detailed correlation table cannot be constructed owing to a paucity of information, it may still be possible to determine the rank of the individual. If proper allowance is made for the necessarily wide inaccuracy of the computed result, the rank coefficient is better than no coefficient at all for such inaccurate or indeterminate data.*

---

* "On Further Methods of Determining Correlation". Drapers Co. Research Memoirs: Bio. Ser. IV.

# CHAPTER XI.

## THE MOMENTS OF A DISTRIBUTION.

**Introduction.** The first moment, obtained by multiplying each deviation by the corresponding frequency, adding the resulting products and dividing by the total frequency of the distribution, was discussed in Chapter IV in connection with the arithmetic mean. The second moment, in which the deviations are squared before multiplication by the frequencies, was discussed in Chapter V. The third and fourth moments, with the deviations cubed and raised to the fourth power respectively, were also referred to in Chapter V.

Obviously the moments may be computed about any point by obtaining the deviations from that point and raising to the appropriate power, etc. For most purposes, however, the second and higher moments are computed about the mean which thus serves as a *standard origin* for the moments.

The moments about the mean are denoted by the symbols $\mu_1$, $\mu_2$,$\mu_3$, $\mu_4$, etc. where the subscripts refer to the *order* of the moments; that is, the index of the power to which the deviations are raised. Under the same system of notation, the moments about any other point are denoted by $\mu_1'$, $\mu_2'$, $\mu_3'$, $\mu_4'$. etc., with the primes serving to distinguish moments about the mean from moments about any other origin.

The moments about the mean may be computed directly by first computing the mean and then subtracting the value of the mean from each deviation and using the resulting differences in the computations for the moments. This method of computing the moments has the advantages of simplicity and directness but it usually leads to troublesome fractions and it ordinarily involves more labor than the indirect methods which are described in this chapter.

**Transformation Formulas for the Moments about the Mean.** The formulas for the moments about the mean in terms of the moments about a fixed point will now be derived.

Let $d$ be the mean deviation, that is, the distance of the mean from the fixed point of reference, and let the $x$'s be measured from the mean. Then corresponding to a given value of $x$ there will be the deviation $x'$ about the fixed point.

From the definition of a moment we have,

$$\mu_1' = \frac{1}{N}\Sigma\,(x+d)y = \frac{1}{N}\Sigma\,xy + \frac{Nd}{N},$$

$$= \mu_1 + d \qquad = d, \text{ since } \Sigma xy \text{ is zero (Theorem, page 35, Chapter IV);}$$

$$\mu_2' = \frac{1}{N}\Sigma\,(x+d)^2y = \frac{1}{N}\Sigma\,x^2y + 2\frac{d}{N}\Sigma\,xy + \frac{Nd}{N}$$

$$= \mu_2 + d^2, \text{ since } \Sigma\,xy = o;$$

$$\mu_3' = \frac{1}{N}\Sigma\,(x+d)^3y = \frac{1}{N}\Sigma\,x^3y + \frac{3d}{N}\Sigma\,x^2y + \frac{3d^2}{N}\Sigma\,xy + \frac{Nd^3}{N}$$

$$= \mu_3 + 3d\mu_2 + d^3;$$

$$\mu_4' = \frac{1}{N}\Sigma(x+d)^4y$$

$$= \frac{1}{N}\Sigma x^4y + \frac{4d}{N}\Sigma x^3y + \frac{6d^2}{N}\Sigma x^2y + \frac{4d^3}{N}\Sigma xy + \frac{Nd^4}{N}$$

$$= \mu_4 + 4d\mu_3 + 6d^2\mu_2 + d^4.$$

Transposing a part of the terms in the four preceding equations and changing the signs, we have the following equations which express each moment about the mean in terms of the corresponding moment about the fixed point and the moments of lower order about the mean:

$$\mu_1 = \mu_1' - d = 0, \text{ since } \mu_1 = d;$$
$$\mu_2 = \mu_2' - d^2;$$
$$\mu_3 = \mu_3' - 3d\,\mu_2 - d^3;$$
$$\mu_4 = \mu_4' - 4d\,\mu_3 - 6d^2\,\mu_2 - d^4.$$

These formulas for transferring the moments from a fixed point to the mean are arranged in what is called the *continuous form;* that is, they begin with the moment of lowest order and proceed step by step up to the fourth moment.

## Exercises.

1. Compute the third and fourth moments for the student height data at the beginning of Chapter III.

2. By taking the fixed point of reference at various points show that for the data of Student Heights the third and fourth moments are least when computed about the arithmetic mean.

3. Show algebraically that any moment is least when taken about the mean.

4. Show that having the formulas arranged in the continuous form does not involve additional labor of computation.

5. By expanding $\mu_n = \frac{1}{N}\Sigma(x'-d)^n$ according to the binominal theorem where $x$ is measured from the fixed point of reference, derive the general formula expressing $\mu_n$ in terms of the moments about the fixed point.

6. Specialize the formula of (5) for the third and fourth moments.

7. Compute the third and fourth moments from the data of Table I by using the formulas of exercise 6.

8. Find the first, second, third and fourth moments about the mean of a distribution with frequencies proportional to the successive terms of the expansion of the binomial $(p+p)^n$.

Ans. $\mu^3 = npq\,(p-q)$; $\mu_4 = npq\,(3(n-2)\,pq+1)$. (See Hardy, "Construction of Tables of Mortality," p. 107 et seq.).

The computation of the moments about the mean either directly or by first computing about a convenient origin and then transforming to the mean is open to the serious practical objection that there are no convenient methods of checking the results. The arithmetic of the following summation methods is comparatively brief and admits of satisfactory checks on the correctness of the results.

**The First Summation Method of Computing the Moments.** The theory of this and the second summation method which follows immediately after it are somewhat detailed but both are entirely elementary throughout.

Let us take a distribution with the five frequencies $y_1, y_2, y_3, y_4, y_5$ corresponding to values of $x$ equal to 1, 2, 3, 4, 5. By the ordinary direct method, the first moment about the point $x = 0$ is $y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5$. Now let us arrange the

$y$'s in vertical order and add in the manner indicated in the second column below.

| (1) | (2) | (3) |
|---|---|---|
| $y_1$ | $y_1 + y_2 + y_3 + y_4 + y_5$ | $y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5$ |
| $y_2$ | $y_2 + y_3 + y_4 + y_5$ | $y_2 + 2y_3 + 3y_4 + 4y_5$ |
| $y_3$ | $y_3 + y_4 + y_5$ | $y_3 + 2y_4 + 3y_5$ |
| $y_4$ | $y_4 + y_5$ | $y_4 + 2y_5$ |
| $y_5$ | $y_5$ | $y_5$ |
| $\Sigma y$ | $y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5$ | $y_1 + 3y_2 + 6y_3 + 10y_4 + 15y_5$ |

| (4) | (5) |
|---|---|
| $y_1 + 3y_2 + 6y_3 + 10y_4 + 15y_5$ | $y_1 + 4y_2 + 10y_3 + 20y_4 + 35y_5$ |
| $y_2 + 3y_3 + 6y_4 + 10y_5$ | $y_2 + 4y_3 + 10y_4 + 20y_5$ |
| $y_3 + 3y_4 + 6y_5$ | $y_3 + 4y_4 + 10y_5$ |
| $y_4 + 3y_5$ | $y_4 + 4y_5$ |
| $y_5$ | $y_5$ |
| $y_1 + 4y_2 + 10y_3 + 20y_4 + 35y_5$ | $y_1 + 5y_2 + 15y_3 + 35y_4 + 70y_5$ |

The sum of the second column is thus the same as the first moment. By the direct method the second moment about the same point is $y_1 + 4y_2 + 9y_3 + 16y_4 + 25y_5$ divided by $N$. Let us designate the sum of column (2), when divided by $N$, by $S$; the second divided by $N$, by $S_2$; the third when divided by $N$, by $S_3$, etc.

That is $S_2 = \dfrac{y_1 + 2y_2 + 3y_3 + \cdots}{N}$, $S_3 = \dfrac{y_1 + 3y_2 + 6y_3 + \cdots}{N}$,

$S_4 = \dfrac{y_1 + 4y_2 + 10y_5 + \cdots}{N}$, $S_5 = \dfrac{y + 5y_2 + 15y_3 + \cdots}{N}$

It is apparent on inspection that $2S_3 - S_2$ is the second moment. In symbols,

$$\frac{2}{N}(y_1 + 3y_2 + 6y_3 + 10y_4 + 15y_5) - \frac{1}{N}(y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5) = \frac{1}{N}(y_1 + 4y_2 + 9y_3 + 16y_4 + 25y_5).$$

That is, $\mu'_2 = 2S_3 - S_2$.

The third moment about the same point of reference is

$$\frac{1}{N}(y_1 + 8y_2 + 27y_3 + 64y_4 + 125y_5).$$

For this moment the following relation is readily verified:

$\mu'_3 = 6S_4 - 6S_3 + S_2$.

Extending the reasoning to the case of the fourth moment, we have

$\mu'_4 = 24S_5 - 36S_4 + 14S_3 - S_2$.

We thus have four relations connecting the moments with the $S$'s:

$$\mu'_1 = S_2,$$
$$\mu'_2 = 2S_3 - S_2,$$
$$\mu'_3 = 6S_4 - 6S_3 + S_2,$$
$$\mu'_4 = 24S_5 - 36S_4 + 14S_3 - S_2.$$

Transferred to the mean as origin by the formulas of page 88 these moments become

$$S_2 = d;$$
$$\mu_2 = \mu'_2 - d^2 = 2S_3 - S_2 - d^2 = 2S_3 - d(1+d);$$
$$\mu_3 = \mu'_3 - 3d\mu_2 - d^3 = 6S_4 - 6S_3 + S_2 - 3d\,\mu_2 - d^2,$$
$$= 6S_4 - 3\mu_2 - 3d(1+d) + d - 3d\,\mu_2 - d^3,$$
$$= 6S_4 - 3\mu_2\,(1+d) - d(1+d)(2+d);$$

and similarly, $\mu_4 = 24S_5 - 2\mu_3\{2(1+d)+1\} - \mu_2\{6(1+d)(2+d) - 1\} - d(1+d)(2+d)(3+d)$.

It is evident that the same relations hold for a larger number of classes than the five which we have assumed for the purpose of illustrating the method.

These relations connecting the moments about the mean with the sums obtained by this process of summation are materially shorter and more convenient than the direct formulas. It will be noticed that the sum of any column is the largest

number in the next column, so that a satisfactory check on the summation is afforded. It is possible, however, by taking the point of reference near the mean, to still further shorten the labor of the computation.

**The Second Summation Method of Computing the Moments.** To illustrate the second method let us take a distribution of eight classes and assume the fixed point of reference at class 5. Then we sum from both top and bottom to, but not including, the frequencies of class 5 in accordance with the following scheme.

| (1) | (2) | (3) | (4) |
|---|---|---|---|
| $y_1$ | $y_1$ | $y_1$ | $y_1$ |
| $y_2$ | $y_2 + y_1$ | $y_2 + 2y_1$ | $y_2 + 3y_1$ |
| $y_3$ | $y_3 + y_2 + y_1$ | $y_3 + 2y_2 + 3y_1$ | $y_3 + 3y_2 + 6y_1$ |
| $y_4$ | $y_4 + y_3 + y_2 + y_1$ | $y_4 + 2y_3 + 3y_2 + 4y_1$ | |
| $y_5$ | | | |
| $y_6$ | $y_6 + y_7 + y_8$ | $y_6 + 2y_7 + 3y_8$ | $y_6 + 3y_7 + 6y_8$ |
| $y_7$ | $y_7 + y_8$ | $y_7 + 2y_8$ | $y_7 + 3y_8$ |
| $y_8$ | $y_8$ | $y_8$ | $y_8$ |

| (5) | (6) |
|---|---|
| $y_1$ | $y_1$ |
| $y_2 + 4y_1$ | |
| $y_6 + 4y_7 + 10y_8$ | $y_6 + 5y_7 + 15y_8$ |
| $y_7 + 4y_8$ | $y_7 + 5y_8$ |
| $y_8$ | $y_8$ |

Forming $\mu'_1$, about the point $x = 5$, by the direct method we have

$$\mu'_1 = \frac{\text{I}}{N}(y_4 + 2y_3 + 3y_2 + 4y_1) + \frac{\text{I}}{N}(y_6 + 2y_7 + 3y_8).$$

But $\mu'_1$ has been defined as equal to $S_2$. Hence $S_2$ is obtained by subtracting the last upper summation term from the last lower term in column (3).

By direct computation,

$$\mu'_2 = \frac{I}{N}(y_4 + 4y_3 + 9y_2 + 16y_1 + y_6 + 4y_7 + 9y_8).$$

But $\mu'_2 = 2S_3 - S_2$, or $2S_3 = \mu'_2 + S_2$.

Hence $2S_3 = \frac{I}{N}(y_4 + 4y_3 + 9y_2 + 16y_1 + y_6 + 4y_7 + 9y_8)$

$$-\frac{I}{N}(y_4 + 2y_3 + 3y_2 + 4y_1 - y_6 - 2y_7 - 3y_8),$$

$$= \frac{I}{N}(2y_3 + 6y_2 + 12y_1 + 2y_6 + 6y_7 + 12y_8).$$

Therefore, $S_3 = \frac{I}{N}(y_3 + 3y_2 + 6y_1) + \frac{I}{N}(y_6 + 3y_7 + 6y_8)$.

That is, $S_3$ is the sum of the last term in the positive, or lower, summation and the last but one (the last term as written in the scheme) of the negative summation terms in column (4).

Likewise,

$S_4 = \frac{I}{N}(y_6 + 4y_7 + 10y_8) - \frac{I}{N}(y_2 + 4y_1)$, the difference between the last positive summation term and the last but two of the negative summation terms in column (5).

And $S_5 = \frac{I}{N}(y_6 + 4y_7 + 15y_8) + \frac{I}{N}y_1$ the sum of the last positive summation term and the last but three of the negative terms in column (6).

After the $S$'s are obtained the formulas of page 91 are applied to obtain the $\mu$'s.

As in the first summation method the partial summations can be added for checks on the Arithmetic.

This second summation method will be found very convenient, especially when the number of classes is large or the frequencies are of considerable size.

The following computations for the data of page 24 illustrate the two summation methods.

*Computations of this length should never be attempted without first arranging a complete form with a place for each number and that place so chosen that the number is in its most convenient location. The entire computation should be planned before the arithmetic is begun.*

| *Class.* | *Freq.* | | | | |
|---|---|---|---|---|---|
| 1 | 2 | 750 | 5925 | 28463 | 105421 |
| 2 | 10 | 748 | 5175 | 22538 | 76958 |
| 3 | 11 | 738 | 4427 | 17363 | 54420 |
| 4 | 38 | 727 | 3689 | 12936 | 37057 |
| 5 | 57 | 689 | 2962 | 9247 | 24121 |
| 6 | 93 | 632 | 2273 | 6285 | 14874 |
| 7 | 106 | 539 | 1641 | 4012 | 8589 |
| 8 | 126 | 433 | 1102 | 2371 | 4577 |
| 9 | 109 | 307 | 669 | 1269 | 2206 |
| 10 | 87 | 198 | 362 | 600 | 937 |
| 11 | 75 | 111 | 164 | 238 | 337 |
| 12 | 23 | 36 | 53 | 74 | 99 |
| 13 | 9 | 13 | 17 | 21 | 25 |
| 14 | 4 | 4 | 4 | 4 | 4 |
| Totals | 750 | 5925 | 28463 | 105421 | 329625 |

$$S_2 = 7.9 \quad S_3 = 37.95 \quad S_4 = 140.56 \quad S_5 = 439.5$$

(1). $d = S_2 = 7.9$

(2). $d(1+d) = 70.31$

(3). $d(1+d)\ (2+d) = 696.069$

(4). $3(1+d) = 26.7$ $\qquad \mu_2 = 2S_3 - d(1+d) = 5.592$

(5). $4(1+d)+2 = 37.6$ $\qquad \sigma = \sqrt{\mu_2} = 2.36$

(6). $6\,(1+d)\,(2+d) - 1 = 527.66$ $\qquad \mu_3 = 6S_4 - \mu_2 .(4). - (3). = -2.015$

$\mu_4 = 24S_5 - \mu_3\,(5) - \mu_2 .(6). - (3+d).(3). = 85.937$

Notice that the sum of the first column = 750, the last sum at the top of the $S_2$ column, and similarly for each following column. The computation form is arranged for the use of a *"Millionaire"* calculating machine.

The second summation method is started as follows:

| *Class* | *Freq.* | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 10 | 12 | 14 | 16 | 18 | 20 |
| 3 | 11 | 23 | 37 | 53 | 71 | 91 |
| 4 | 38 | 61 | 98 | 151 | 222 | |
| 5 | 57 | 118 | 216 | 367 | | |
| 6 | 93 | 211 | 427 | | | |
| 7 | 106 | | | | | |
| 8 | 126 | 433 | 1102 | 2271 | 4477 | 8085 |
| 9 | 109 | 307 | 669 | 1269 | 2206 | 3608 |
| 10 | 87 | 198 | 362 | 600 | 937 | 1402 |
| 11 | 75 | 111 | 164 | 238 | 337 | 465 |
| 12 | 23 | 36 | 53 | 74 | 99 | 128 |
| 13 | 9 | 13 | 17 | 21 | 25 | 29 |
| 14 | 4 | 4 | 4 | 4 | 4 | 4 |
| Totals..... | 750 | 1102 | 2271 | 4477 | 8085 | |

$$S_2 = (1102 - 427)/750 = 0.9$$
$$S_3 = (2271 + 367)/750 = 3.52$$
$$S_4 = (4477 - 222)/750 = 5.67$$
$$S_5 = (8085 + 91)/750 = 10.9$$

The computation from this point on is the same as under the first method, except that the origin is at class 7, or the height class 67, instead of height class 60.

## Exercises.

9. Compute the moments for the frequency distribution of page 29 by the two summation methods.

10. Demonstrate the proof of the two summation methods for $n$ classes.

11. What difference would result in the computations of the second summation method if the origin were taken at the eighth class instead of the seventh so that the upper sum in the first summation is the larger?

**Correction Formulas for the Moments.** All the methods that have been proposed for finding the moments assume that the frequencies are concentrated at the center of each class while actually the deviations are continuously distributed from one end of the range to the other so that there is nothing in the nature of the data to correspond to the classes, mid-ordinates, etc. A certain degree of error is therefore introduced by these methods. We are not really working with the actual deviations but with the artificial classes built up from the actual deviations. In how far then are facts, which hold for the classes, of significance for the actual variates? It may well be that in ordinary statis-

tical work the closeness of the measurements may not warrant taking these errors into account but the corrections are easily applied and frequently make a significant difference in the results. However the corrections should not be applied to data not accurate enough to warrant such care no matter if the corrections are easily applied. ***The methods adopted in computation must never be such as to presuppose more accurate data than that in hand.***

When the distinction is made between the moments as calculated from the class frequencies and deviations and the moments calculated under the assumption of continuous variation, it is customary to denote the values as computed by $\nu_1$, $\nu_2$, $\nu_3$, $\nu_4$, and $\nu_1'$, $\nu_2'$, $\nu_3'$, $\nu_4'$, and reserve the corresponding $\mu$'s for the values under the assumption of continuity. When no account is taken of the distinction between the discrete and continuous series of frequencies, the $\mu$'s alone are used. The $\nu$'s are often spoken of as the **raw** or **unadjusted moments** and the $\mu$'s as the **adjusted** moments.

The adjustment or correction formulas are:

$$\mu_1 = \nu_1 = 0$$
$$\mu_2 = \nu_2 - \tfrac{1}{12}$$
$$\mu_3 = \nu_3$$
$$\mu_4 = \nu_4 - \tfrac{1}{2}\nu_3 + \tfrac{7}{240}$$

The theory of these corrections is due to Dr. Sheppard and to Professor Pearson. A simple demonstration of the formulas is that of Bio. III, p. 308.

According to the underlying mathematical theory these correction formulas hold in strictness only for a frequency curve with high contact at each end. When these conditions are not satisfied it is probably best not to apply the corrections.

**Theorem I.** *Changing the unit of measurement of the deviation; that is, multiplying each deviation by a constant, multiplies a moment by that constant raised to a power equal to the order of the moment.* For,

$$\mu_n = \frac{1}{N}\Sigma x^n y$$

and $\Sigma(rx)^n y = r^n \Sigma x^n y$.

**Theorem II.** *Multiplying or dividing each frequency by a constant does not change the moments.* For,

$$\frac{\Sigma x^n r y}{\Sigma r y} = \frac{r\Sigma x^n y}{r\Sigma y} = \frac{\Sigma x^n y}{\Sigma y}.$$

Because the values of the third and fourth moments depend on the unit of measure of the deviations it is usual to employ these two moments in the forms $\beta_1$ and $\beta_2$, respectively, where $\beta_1 = \mu_3^2/\mu_2^3$ and $\beta_2 = \mu_4/\mu_2^2$. To show that $\beta_1$ and $\beta_2$ are independent of the unit of measure of $x$ let us write $\beta_1 = \frac{N(\Sigma x^3 y)^2}{(\Sigma x^2 y)^3}$ and $\beta_2 = \frac{N(\Sigma x^4 y)}{(\Sigma x^2 y)^2}$. Then let $x$ be changed into $rx$ where $r$ is any constant.

This gives $\beta_1 = \frac{N(\Sigma x^3 y)^2 . r^6}{(\Sigma x^2 y)^3 . r^6} = \frac{N(\Sigma x^3 y)^2}{(\Sigma x^2 y)^3}$, and similarly for $\beta^2$.

## Exercises.

12. Show that adding a constant to each deviation changes the moments.

13. Show that adding a constant to each frequency changes the moments.

**Summations.** The following exercises are intended for practice in using summations and should be carefully worked through in order that a comprehension of the somewhat detailed discussions of subsequent chapters is not hindered by a lack of familiarity with the necessary algebra.

## Exercises.

14. Show that the square of $\Sigma xy$ is $\Sigma x^2 y^2 + \Sigma\Sigma x_s y_s x_t y_t$ where the subscripts are attached in the second summation to indicate the product of unequal deviations, and all deviations are measured from the mean.

15. By actually computing the separate value of each summation verify the relation $(\Sigma xy)^2 = \Sigma x^2 y^2 + 2\Sigma\Sigma x_s y_s x_t y_t$ for the distribution 1, 2, 5, 2, 1.

16. Establish the relation $(\Sigma xy)^3 = \Sigma x^3 y^3 + 3\Sigma\Sigma x_s^2 y_s^2 x_t y_t$.

17. Establish the relation $(\Sigma xy)^4 = \Sigma x^4 y^4 + 4\Sigma\Sigma x_s^3 y_s^3 x_t y_t + 6\Sigma\Sigma x_s^2 y_s^2 x_t^2 y_t^2$.

18. Show that $(\Sigma xy)(\Sigma x^2 y) = \Sigma x^3 y^2 + \Sigma\Sigma x_s y_s x_t^2 y_t$.

19. Prove that $x_s^2 + x_t^2 > 2x_s x_t$ and hence $\Sigma(x_s^2 + x_t^2) > 2\Sigma x_s x_t$.

20. Show that $(\Sigma x^4 y)(\Sigma y) > \Sigma (x^2 y)^2$.

We have $\Sigma x^4 y = x^4_1 y_1 + x^4_2 y_2 + x^4_3 y_3 + \ldots$

and $(\Sigma y)(\Sigma x^4 y) = x^4_1 y^2_1 + x^4_2 y^2_2 + x^4_3 y^2_3 + \ldots + x^4_1 y_1 y_2 + x^4_1 y_1 y_3 + \ldots + x^4_2 y_2 y_1 + x^4_2 y_2 y_3 + \ldots + \ldots$

$= \Sigma x^4 y^2 + \Sigma x^4_s y_s y_t = \Sigma x^4 y^2 + \Sigma y_s y_t (x^4_s + x^4_t)$,

Also $(\Sigma x^2 y)^2 = \Sigma x^4 y^2 + 2\Sigma x^2_s y_s y_t = \Sigma x^4 y^2 + 2\Sigma y_s y_t x^2_s x^2_t$.

Therefore $(\Sigma x^4y)\ (\Sigma y) > (\Sigma x^3y)^2$,
if $\Sigma x^4y^2 + \Sigma y_s y_t (x^4_s + x^4_t) > \Sigma x^4y^2 + 2\Sigma y_s y_t x^3_s x^3_t$.
i. e. if $\Sigma y_s y_t (x^4_s + x^4_t) > 2\Sigma y_s y_t x^3_s x^3_t$.

But the sum of the squares of two quantities is always greater than twice their products and hence each term on the left is greater than the corresponding term on the right, thus proving the theorem. The algebraic discussion may be more easily followed if a summation of only two or three terms is first employed.

21. Prove that $(\Sigma x^4y)\ (\Sigma x^2y) > (\Sigma x^3y)^2$.

22. Prove that $\beta_2 > \beta_1$.

**The Moments and the Equation of the Smoothed Curve.** It is shown in Chapter II that a smooth curve is fitted on the basis of *principles* which are assumed true for the data as a whole. One such principle is that of *equality of area* which assumes that the area under the curve is equal in numerical value to the total frequency of the distribution.

The **principle of equality of moments** assumes in addition to the equality of area and total frequency that the first, second, third and fourth moments computed directly from the data are respectively equal to the first, second, third and fourth moments computed from the adjusted frequencies.

To illustrate the application of the method of equality of moments let us fit a straight line to the points (2,4), (3,3), (4,7), (5,6).

The equation of the required line is $y = mx + b$ where $m$ and $b$ are to be determined. The adjusted $y$'s in terms of $m$ and $b$ are $2m + b$, $3m + b$, $4m + b$, $5m + b$.

The equality of the area and the total frequency can be expressed as an equality of moments if the moment of zero order is permitted. This is possible because any number with an exponent zero is equal to unity. Hence $2^\circ \cdot 4 + 3^\circ \cdot 3 + 4^\circ \cdot 7 + 5^\circ \cdot 6 = 4 + 3 + 7 + 6 = 20$. Also, $2^\circ \cdot (2m + b) + 3^\circ \cdot (3m + b) + 4^\circ \cdot (4m + b) + 5^\circ \cdot (5m + b) = 14m + 4b$.

Hence, on equating the two zero moments,

$$14m + b = 20.$$

From the first moment, $2.4 + 3.3 + 4.7 + 5.6 = 2(2m + b) + 3(3m + b) + 4(4m + b) + 5(5m + b)$, we have

$$51m + 14b = 75.$$

Solving these two moment equations simultaneously we have

$$m = 2.5$$
$$b = -3.75$$

Therefore $y = 2.5x - 3.75$ is the required equation of the straight line fitted to the given points on the basis of the asumption of equality of the zero and first moments respectively.

## Exercises.

23. Fit a straight line to the preceding illustrative points on the assumption of equality of the first and second moments respectively. Should the resulting equation agree exactly with the equation found above?

24. Fit a straight line to the points (1,5), (3,8), (4,6), (5,5), (7,10).

25. Fit a parabola, $y = a + bx + cx^2$, to the points of Exercise 24.

## CHAPTER XII.

### FURTHER THEORY OF CORRELATION.

**A Second Concept of Correlation.** In Chapter VII two attributes are said to be correlated when there is a tendency for a change in the value of one to be followed by a change in the value of the other. And the ratio of the standard deviation of the means of the arrays to the standard deviation of all the variates was taken as the measure of the degree of correlation between the attributes. A second approach to the matter of correlated variates is as follows.

On the assumption *that the mean is the representative of the variates of an array* the dependence of $y$ on $x$ is exhibited by the curve of means; that is, by the regression curve. Obviously this curve is a significant measure of the dependence of $y$ on $x$ only insofar as the means are *in fact* representatives of the variates of the respective arrays. Within this limitation the spread of the variates about the means of the successive arrays is a measure of the extent of dependence of $y$ on $x$; that is, of the correlation of $y$ with $x$.

Let $\sigma_{a_y}$ denote the mean squared divergence from the regression curve. Then

$$N\sigma^2_{a_y} = \Sigma\Sigma n_{xy}(y - \bar{y}_y)^2.$$

As is explained in Chapters VIII and IX, this mean squared deviation must be divided by $\sigma_y^2$ in order to obtain a correlation index of value for purposes of comparison.

To reduce $\sigma_{a_y}^2$ we write,

$$\begin{aligned}\Sigma\Sigma n_{xy}(y - \bar{y}_x)^2 &= \Sigma\Sigma n_{xy}(y - \bar{y} + \bar{y} - \bar{y}_x)^2,\\ &= \Sigma\Sigma n_{xy}(y - \bar{y})^2\\ &+ 2\Sigma\Sigma n_{xy}(y - \bar{y})(\bar{y} - \bar{y}_x)\\ &+ \Sigma\Sigma n_{xy}(\bar{y} - \bar{y}_x)^2,\\ &= N\sigma_y^2 + N\eta^2\sigma_y^2\\ &+ 2\Sigma\Sigma n_{xy}(y - \bar{y})(\bar{y} - \bar{y}_x).\end{aligned}$$

$$\text{But } 2\Sigma\Sigma n_{xy}(y-\bar{y})(\bar{y}-\bar{y}_x) = 2\Sigma_x(\bar{y}-\bar{y}_x)\Sigma_y n_{xy}(y-\bar{y})$$
$$= 2\Sigma_x(\bar{y}-\bar{y}_x)n_x(\bar{y}_x-\bar{y})$$
$$= -2\Sigma n_x(\bar{y}_x-\bar{y})^2$$
$$= -2N\eta^2\sigma_y^2.$$

Therefore $N\sigma_y^2 = N\sigma_y^2 - N\eta^2\sigma_y^2$, and $\dfrac{\sigma_{a_y}^2}{\sigma_y^2} = 1-\eta^2$.

On division by $N\sigma_y^2$, we have,

$$\frac{\sigma_{a_y}^2}{\sigma_y^2} = (1-\eta^2).$$

It accordingly appears that this second measure of correlation is not independent of the first, being equal to $1-\eta^2$.

## Exercises.

1. Show that the mean spread of the variates about the regression line of Chapter IX. is equal to $(1-r^2)\sigma_y^2$.

2. Show that $\sigma_{a_y}^2$ is zero for perfect correlation and equal to $\sigma_y^2$ for zero correlation.

3. Discuss the convenience of the computation formula $\eta^2 = 1 - \sigma_{a_y}^2/\sigma_y^2$.

4. Which approach to the numerical measure of correlation seems clearer?

**Derivation of the Equations of the Regression Lines.** Let the regression equations be of the form:

$$\bar{y}_x = b_{yx}\,.\,x + a.$$
and $$\bar{x}_y = b_{xy}\,.\,y + c.$$

The moments of order zero and unity give for the regression of $y$ on $x$ two moment equations:

$$\Sigma n_x\bar{y}_x = b_{yx}\Sigma n_x\,.\,x + Na,$$
and $$\Sigma n_x\bar{y}_x\,.\,x = b_{yx}\Sigma n_x\,.\,x^2 + a\Sigma n_x\,.\,x.$$

As explained on page 75 the moments are computed for each individual frequency of an array. Hence we have $\Sigma n_y\,.\,\bar{y}_x$ and not merely $\Sigma\bar{y}_x$. For the same reason $a$ appears in the moment sum once for each frequency; that is, $N$ times.

Since $\bar{y}_x = \frac{1}{n_x} \Sigma_y n_{xy} \cdot y$, $\Sigma_x n_x \bar{y}_x = \Sigma_x n_x \cdot \frac{1}{n_x} \Sigma_y n_{xy} \cdot y =$

$= \Sigma_x \Sigma_y n_{xy} y$, a first moment about a horizontal line thru the mean, (both $y$ and $x$ are assumed to be measured from their respective means) and hence zero from an obvious extension of the theorem of page 35.

Likewise $\Sigma n_x \cdot x = o$. Hence, from the first moment equation, a as equal to zero.

In the second equation we have,

$$\Sigma n_x \bar{y}_x \cdot x = \Sigma\Sigma n_{xy} \cdot y \cdot x.$$

The summation $\Sigma n_x \cdot x^2$ has been taken equal to $N\sigma_x^2$ and $\Sigma n_y \cdot y^2$ equal to $N\sigma_y^2$. It seems consistent with this notation to assume $\Sigma\Sigma n_{xy} yx = Nr\sigma_x\sigma_y$ where $r$ is the numerical constant of Chapter IX.

On reducing the second moment equation we have

$$N r\sigma_x\sigma_y = b_{yx} \cdot N\sigma_x^2$$

Therefore $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$

and hence $\bar{y}_x = r \cdot \frac{\sigma_y}{\sigma_x} \cdot x$ is the required regression equation.

## Exercises.

5. Derive the regression equation $\bar{x}_y = r \frac{\sigma_x}{\sigma_y} y$.

6. Prove in detail that $\Sigma\Sigma n_{xy} y = o$, where $x$ and $y$ are measured from the mean.

When $x$ and $y$ are measured from the original axes the regression equations become

$$\bar{y}_x - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$\bar{x}_y - \bar{x} = r \frac{\sigma_x}{\sigma_y} (y - \bar{y}).$$

**The Relation Between $\eta$ and r.** It was shown in Chapter IX that $\eta$ and $r$ have the same numerical value when the regres-

sion is truly linear. Hence a lack of agreement in the values of $\eta$ and $r$ is an indication of a divergence from linearity in the regression. The difference between $\eta$ and $r$ is expressible in terms of the divergence from linearity by the two equations:

$$N\sigma_y^2(\eta_y^2 - r^2) = \Sigma n_x(\bar{Y}_x - \bar{y}_x)^2$$

and $N\sigma_x^2(\eta_x^2 - r^2) = \Sigma n_y(\bar{X}_y - x_y)^2$, where $\bar{Y}_x$ and $\bar{X}_x$ are the regression line means.

To prove the first of these formulas let us add and subtract $\bar{y}$ for each term in the summation $\Sigma n_x(\bar{Y}_x - \bar{y}_x)^2$. We then have after expansion,

$$\Sigma n_x(\bar{Y}_x - \bar{y}_x)^2 = \Sigma n_x\{(\bar{Y}_x - \bar{y})^2 - 2(\bar{Y}_x - \bar{y})(\bar{y}_x - \bar{y}) + (\bar{y}_x - \bar{y})^2\}.$$

On substituting from the regression equations this expanded form becomes

$$\Sigma n_x(x - \bar{x})^2 \cdot r^2\frac{\sigma_y^2}{\sigma_x^2} + \Sigma n_x(\bar{y}_x - \bar{y})^2 - 2\Sigma n_x \cdot r\frac{\sigma_y}{\sigma_x}(\bar{y}_x - \bar{y})(x - \bar{x}),$$

which equals

$$N\sigma_x^2 \cdot r^2\frac{\sigma_y^2}{\sigma_x^2} + N\sigma_y^2 \cdot \eta_y^2 - 2r\frac{\sigma_y}{\sigma_x}Nr\sigma_x\sigma_y.$$

Substituting for the last term and collecting we have

$$\Sigma n_x(\bar{Y}_x - \bar{y}_x)^2 = N\sigma_y^2(\eta_y^2 - r^2).$$

## Exercises.

7. Prove the formula $\Sigma n_y(\bar{X} - \bar{x}_y) = N\sigma_x^2(\eta_x^2 - r^2)$.
8. Show from these formulas that $\eta > r$.
9. Show that the same pair of equations will be obtained for the regression lines if the assumes lines are fitted to the individual frequencies instead of to the means of the arrays.
10. Prove that for truly linear regression $\sigma n_y = r\sigma_y$.
11. Show that for truly inear regression $\sigma x_y^2 = \sigma_y^2(1 - r^2)$.

**The Coefficient r for Non-linear Regression.** It has been seen on page 85 that $r$ is always too small for any but strictly linear regressions. This is due to the fact that the summation $\Sigma\Sigma n_{xy}(x - \bar{x})(y - \bar{y})$ involves both positive and negative terms which cancel each other with a consequent reduction in the value

of the summation. If the regression curve is carefully drawn a fair idea of the trustworthiness of $r$ can be obtained by observing the departures of that curve from linearity. A more accurate way, of course, is to compute both $\eta$ and $r$ and observe the difference in value of the two constants.

$$\text{Since} \quad r = \frac{\Sigma\Sigma n_{xy}(x - \bar{x})(y - \bar{y})}{N\sigma_x\sigma_y}$$

the size of $r$ varies directly as the value of the summation in the numerator. In this summation the large values of both $x$ and $y$ are along either diagonal and hence $r$ will be largest numerically when the values of $n_{xy}$ are largest along a diagonal. If the frequencies tend to lie along one diagonal the value of $r$ will be positive; along the other, negative. If the distribution should exhibit two tendencies, — to concentrate along both diagonals — the cancellation of terms with opposite signs would give rise to a small value for $r$. Also the regression may be markedly nonlinear, circular, or periodic as a sine curve, so that the straight line fitted to the means of the arrays is practically horizontal, resulting in a very small value for $r$. This may be true even for data which shows a definite tendency for the frequencies to cluster closely along the curve of means; that is, it is possible for $r$ to have a small value even though the data shows the attributes to have in fact a high degree of correlation. In any of these or similar cases the correlation should be determined from the *correlation ratio* which is not affected by the form of the regression.

**The Most Probable Value of a Characteristic** can be determined from $r$. Let us first define the properties, *homoscedasticity* and *homoclisy*.

The mean standard deviation of the frequencies of an array has been denoted by the symbol $\sigma_{a_y}^2$ where $\sigma_{a_y}^2 = \sigma_y^2(1 - \eta^2)$ or in terms of $r$, $\sigma_{a_y}^2 = \sigma_y^2(1 - r^2)$. It must be remembered that these are mean values so that it may well happen that the true standard deviation of an individual array may differ considerably from these values. A distribution in which all arrays of a given sense, that is, all $y$ or all $x$ arrays have the same standard deviation is said to be **homoscedastic** with respect to the arrays of that sense.

It has been assumed that the frequencies of the arrays are so distributed that the means and the modes coincide; that is, so that the mean is the most probable value of the array, but this may not always be even approximately true. The arrays of a distribution are said to be **homoclitic** when the mean is the most probable value of the array.

On the basis of the just preceding definitions it may be said that for homoclitic arrays the most probable value of $y$ corresponding to a given value for $x$ is found from the equation

$$y = r \frac{\sigma_y}{\sigma_x} \bar{x}, \text{ or}$$

$$y - \bar{y} = r \frac{\sigma_y}{\sigma_x} (x - \bar{x}).$$

A knowledge of the most probable values is of little importance unless accompanied by information of the dispersion about that value; that is, of the standard deviation and the probable deviation. Since the entire theory of estimating values of a characteristic is based on the coefficient of correlation the probable deviation of $y$ when obtained from the regression curve is logically $0.67459\, \sigma_y \sqrt{(1 - r^2)}$, and not $0.67449\, \sigma_y \sqrt{(1 - \eta^2)}$ (provided the arrays are homoscedastic, otherwise no general formula is possible and the dispersion of each array must be computed directly from the data of the respective arrays). Likewise the probable error of $x$ found from the regressions is $0.67459\, \sigma_x \sqrt{(1 - r^2)}$, with the same restrictions as to homoscedasticity.

If the three conditions of linearity of regression, of homoscedasticity, and of homoclisy are satisfied the just preceding theory of estimating the value of a variable characteristic is complete and practically valuable. In ordinary distributions these conditions are likely to hold at least approximately so that when intelligently applied the theory is of importance. In every case the regression curve should be determined graphically and both $\eta$ and $r$ computed and the difference in their values noted, and the test for linearity applied. If there is doubt as to the homoscedasticity, the standard deviations can be computed directly from the arrays in question and the probable deviations determined from the resulting values instead of from the pre-

ceding formula. The question of homoclisy is usually disregarded though wide departures should be noted and taken into consideration.

## Exercises.

12.* What is the most probable weight of a student of height 70 inches?

13. What is the most probable height of a student of weight 132 pounds?

14. What is the most probable rainfall of a month with a mean temperature of 54 degrees?

15. What is the most probable top beef cattle price for a month with a top hog price of $8.25?

16. Compute the probable deviations from the most probable values of Exercises 5, 6, 7, and 8.

17. Discuss the practical reliability of the preceding estimates. In how far is the probable deviation a trustworthy index of this reliability?

If two **distributions are superimposed** the value of r for the combined group is connected with the constituent r's by the following relation.

$$N\sigma_y\sigma_x r = N_1\sigma_{y_1}\sigma_{x_1}r_1 + N_2\sigma_{y_2}\sigma_{x_2}r_2 + \frac{N_1 N_2}{N}(\bar{x}_1 - \bar{x}_2)\ (\bar{y}_1 - \bar{y}_2).$$

The proof of this equation is left as an exercise.

Two superimposed distributions with the same means have the simple relation of connection with the r's:

$$N\sigma_y\sigma_x r = N_1\sigma_{y_1}\sigma_{x_1}r_1 + N_2\sigma_{y_2}\sigma_{x_2}r_2.$$

form which the effect of various mixtures of data can be readily traced. For instance, if the second distribution has a constant frequency for each subgroup, $r_2$ is zero and the value of $r$ is smaller than that of $r_1$ in the proportion $\frac{N_1\sigma_{y_1}\sigma_{x_1}}{N_1\sigma_{y_2}\sigma_{x_1}}$. *That is, adding a constant to each frequency decreases the value of r.*

The **effect of multiplying each frequency** is readily determined.

Let $n_x$ be replaced by $an_x$ and $n_y$ by $an_y$.

$$r = \frac{\frac{\Sigma\Sigma n_{xy} \cdot x \cdot y \cdot}{N}}{\frac{\Sigma n_x x^2}{N} \cdot \frac{\Sigma n_y y^2}{N}}$$ becomes

* Exercises 12, 13, 14, 15, refer to data already given.

$$r = \frac{\dfrac{a\Sigma\Sigma n_{xy} \cdot x \cdot y \cdot}{aN}}{\dfrac{a\Sigma n_x \cdot x^2}{aN} \cdot \dfrac{a\Sigma n_y y^2}{aN}}$$

$$= \frac{\Sigma\Sigma n_{xy} \cdot x \cdot y}{\Sigma n_x \cdot x^2 \cdot \Sigma n_y \cdot y^2} \text{ as before.}$$

That is, *multiplying each frequency is without effect on the value of r.*

**Correlation of indices.** A mathematical measure of spurious correlation will now be derived.

Let us take the two series of measurements $x$ and $y$ and let $I_{xy} = x/y$ be the corresponding series of indices.

The mean $\bar{I}$ *cannot, except under certain conditions, be obtained by dividing the mean of the x's by the mean of the y's.* For, by definition

$$\bar{I}_{xy} = \frac{1}{N} \Sigma \frac{x}{y}$$

Transferring to the respective means, $x$ and $y$, as origins we have

$$\bar{I}_{xy} = \frac{1}{N} \Sigma \frac{\bar{x} + \delta x}{\bar{y} + \delta y}, \text{ where the } \delta\text{'s denote merely the new variables}$$

measured from the respective means.

On rearranging

$$\bar{I}_{xy} = \frac{1}{N} \Sigma \left( \frac{1 + \dfrac{\delta x}{\bar{x}}}{1 + \dfrac{\delta y}{\bar{y}}} \cdot \frac{\bar{x}}{\bar{y}} \right),$$

$$= \frac{1}{N} \cdot \frac{\bar{x}}{\bar{y}} \Sigma \left(1 + \frac{\delta x}{\bar{x}}\right) \left(1 + \frac{\delta y}{\bar{y}}\right)^{-1},$$

$$= \frac{1}{N} \cdot \frac{\bar{x}}{\bar{y}} \Sigma \left\{ 1 + \frac{\delta x}{\bar{x}} - \frac{\delta y}{\bar{y}} - \frac{\delta x \delta y}{\bar{x}\bar{y}} \right\},$$

as far as second terms.

But $\Sigma\delta x = \Sigma\delta y = o$; $\delta x \delta y = N r \sigma x \sigma y$; $\Sigma\delta y = N\sigma_y^2$ and $\Sigma 1 = N$.

$$\text{Hence, } \bar{I}_{xy} = \frac{\bar{x}}{\bar{y}}\left(1 + \frac{\sigma_y^2}{\bar{y}^2} - \frac{r_{xy}\sigma_x\sigma_y}{\bar{x}\,\bar{y}}\right)$$

$$= I_{\bar{x}\bar{y}}\left(1 + \frac{\sigma_y^2}{\bar{y}^2} - \frac{r_{xy}\sigma_x\sigma_y}{\bar{x}\,\bar{y}}\right)$$

This formula shows that the error in assuming $\bar{I}_{xy} = I_{\bar{x}\bar{y}}$ may be considerable. It is smallest where the data is perfectly correlated and largest for uncorrelated material.

## Exercises.

18. Let indices be formed from the two series of measurements

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1 | 3 | 4 | 3 | 2 | 7 | 4 |

Find the value of $\bar{I}_{xy}$ and compare with $I_{\bar{x}\bar{y}}$.

**The Standard Deviations** $\sigma_I$ of the index $\bar{I}_{yx}$ is given by the equation

$$\sigma^2_{\bar{I}_{xy}} = \bar{I}_{xy}^2\left(\frac{\sigma x^2}{\bar{x}^2} + \frac{\sigma y^2}{\bar{y}^2} - 2r\frac{\sigma x}{\bar{x}}\cdot\frac{\sigma y}{\bar{y}}\right).$$

The standard deviation will be computed about the corrected mean. We have

$$N\sigma_I^2 = \Sigma\left(\frac{x}{y} - \bar{I}\right)^2$$

$$= \Sigma\left\{\frac{\bar{x} + \delta x}{\bar{y} + \delta y} - \frac{\bar{x}}{\bar{y}}\left(1 - \frac{r\sigma_x\sigma_y}{\bar{x}\bar{y}} - \frac{\sigma_y^2}{\bar{y}^2}\right)\right\}^2,$$

On dropping the cross term,

$$= I^2_{\bar{x}\bar{y}}\Sigma\left\{\left(1 + \frac{\delta x}{\bar{x}}\right)\left(1 + \frac{\delta y}{\bar{y}}\right)^{-1} - \left(1 - \frac{r\sigma_x\sigma_y}{\bar{x}\bar{y}} + \frac{\sigma_y^2}{\bar{y}^2}\right)\right\}^2$$

$$= I^2_{\bar{x}\bar{y}}\Sigma\left(\frac{\sigma_x}{\bar{x}} - \frac{\sigma_y}{\bar{y}} + \text{squared terms}\right)^2$$

$$= I^2_{\bar{x}\bar{y}}\left(\frac{\sigma x^2}{\bar{x}^2} + \frac{\sigma y}{\bar{y}^2} - 2r\frac{\sigma x}{\bar{x}}\frac{\sigma y}{\bar{y}}\right)$$

## Exercises.

19. Prove that exactly the same formula is obtained for the standard deviation about the unadjusted mean $\frac{\bar{x}}{\bar{y}}$.

The theorem just stated as an exercise shows that in so far as the standard deviation of a distribution of index number is concerned, it is immaterial whether the index of the means, $\bar{x}/\bar{y}$ is corrected or not.

By a method very similar to that employed in the case of the standard deviation, it may be shown that the coefficient of correlation between indices is given by the formula

$$r_{I:I} = \frac{\frac{\sigma_x}{\bar{x}}\frac{\sigma_z}{\bar{z}} r_{xz} - \frac{\sigma_x}{\bar{x}}\frac{\sigma_w}{\bar{w}} r_{xw} - \frac{\sigma_y}{\bar{y}}\frac{\sigma_z}{\bar{z}} r_{yz} + \frac{\sigma_y}{\bar{y}}\frac{\sigma_w}{\bar{w}} r_{yw}}{\sqrt{\frac{\sigma_x^2}{\bar{x}^2} + \frac{\sigma_y^2}{\bar{y}^2} - 2r_{xy}\frac{\sigma_x}{\bar{x}}\frac{\sigma_y}{\bar{y}}}\sqrt{\frac{\sigma_z^2}{\bar{z}^2} + \frac{\sigma_w^2}{\bar{w}^2} - 2r_{zw}\frac{\sigma_z}{\bar{z}}\frac{\sigma_w}{\bar{w}}}}$$

When the four variables $x$, $y$, $z$ and $w$ are uncorrelated the value of $r_{I:I}$ is zero. When the bases are constants $\sigma_y = \sigma_w = 0$, and we have

$$r_{I:I} = \frac{\frac{\sigma_x}{\bar{x}}\cdot\frac{\sigma_z}{\bar{z}}}{\frac{\sigma_x}{\bar{x}}\cdot\frac{\sigma_z}{\bar{z}}} \cdot r_{xz} = r_{xz}$$

That is, dividing each value of a characteristic by the same constant does not affect the value of the coefficient of correlation.

As a special case of this result, *if the absolute values of two characteristics are correlated the degree of correlation is not changed by expressing the measurements as percents.*

When the bases $y$ and $w$ are identical $r_{I:I}$ takes the value,

$$r_{I:I} = \frac{\frac{\sigma_x}{\bar{x}}\frac{\sigma_z}{\bar{z}} r_{xz} - \frac{\sigma_x}{\bar{x}}\frac{\sigma_y}{\bar{y}} r_{xy} - \frac{\sigma_y}{\bar{y}}\frac{\sigma_z}{\bar{z}} r_{yz} + \frac{\sigma_y^2}{\bar{y}^2}}{\sqrt{\left\{\frac{\sigma_x^2}{\bar{x}^2} + \frac{\sigma_z^2}{\bar{z}^2} - 2\frac{\sigma_x}{\bar{x}}\frac{\sigma_z}{\bar{z}} r_{xz}\right\}}\sqrt{\left\{\frac{\sigma_y^2}{\bar{y}^2} + \frac{\sigma_z^2}{\bar{z}^2} - 2\frac{\sigma_z}{\bar{z}}\frac{\sigma_y}{\bar{y}} r_{yz}\right\}}}$$

Now let $x$, $y$ and $z$ be entirely uncorrelated. Then

$$r_{I:I} = \frac{\frac{\sigma_y^2}{\bar{y}^2}}{\sqrt{\left(\frac{\sigma_x^2}{\bar{x}^2} + \frac{\sigma_y^2}{\bar{y}^2}\right)}\sqrt{\frac{\sigma_z^2}{\bar{z}^2} + \frac{\sigma_y^2}{\bar{y}^2}}}$$

This last value of $r_{I:I}$ may even equal 0. 5, in the case when

$$\frac{\sigma_y}{y} = \frac{\sigma_x}{x} = \frac{\sigma_z}{z}.$$

*Hence by dividing each deviation by a third variable, it is possible to introduce correlation into strictly uncorrelated material to as great an extent as 0.5.*

Care must therefore be taken in dealing with index numbers that the full value of $r$ is significant for the absolute values of the measurements. By computing r from the formula

$$r = \frac{\frac{\sigma_y^2}{\bar{y}^2}}{\sqrt{\frac{\sigma_x^2}{\bar{x}} + \frac{\sigma_y^2}{\bar{y}^2}}\sqrt{\frac{\sigma_z^2}{\bar{z}^2} + \frac{\sigma_y^2}{\bar{y}^2}}}$$

the value of the greatest possible degree of spurious correlation is obtained. A value of $r$ greater than this value is certainly significant; a value less may be significant but must be accepted with caution.

*Since by the formula the spurious correlation is zero when the standard deviation or variability of y is zero, it follows that the base of a system of index numbers should be as nearly constant as possible.*

A theory of spurious correlation might be developed for the correlation ratio but the algebraiac details are so much more workable for the correlation coefficient that it would be hardly worth the extra effort. It is conceivable that such a theory would be practically necessary but it is unlikely because after all only approximate results are valuable. There would be little of value in attempting to measure the degree of spurious correlation with precision.

It must be remembered that the matter of spurious correlation is essentially one of interpretation. The question is what does correlation mean. *The correlation is actual and real for the indices but it may be spurious in so for as the absolute values of the measurements are concerned.*

## CHAPTER XIII.

### THE METHOD OF CONTINGENCY.

### THE CORRELATION OF NON-QUANTITATIVE CHARACTERISTICS.

**The Mean Squared Contingency.** — Both the correlation ratio and the correlation coefficient are based on the variation in the means of the arrays; the first directly and the second through the straight line fitted to the points located by the means. Another method of measuring the degre of connection between the attributes, described and illustrated in this chapter, is based on elementary notions of probability. It may be stated in beginning the discussion of the method of contingency that not the least important value of a study of the subject is the additional insight which it gives into the fundamental nature of correlation.

In Table I the distribution of height without reference to weight is given by the total frequency of the $y$ arrays; that is, by the totals at the foot of the table, and the distribution of weight without reference to the distribution of height by the column of sums at the right in the table.

When there is no tendency for certain weights to be most often associated with certain heights, the frequency of a subgroup should be proportional to the total frequencies of its two arrays. Thus imagine the frequencies of the subgroups erased from Table VIII of Chapter VII and then filled in entirely at random; that is, without bias or selection. Since 110/750 of the total frequency of the distribution appears in the height array of weight type 137; that is, since 110 individuals out of 750 are of weight 137, it is logical to assume that this height array contains 114/750 of the frequency of each array which it crosses The frequency of the subgroup (68,-137), for instance, should be 110/750 of 126. *And in general, when the individuals are placed at random, the frequency of a subgroup is given by the formula* $\frac{n_x \cdot n_y}{N}$. For the $y$ array of

type $x$ contains $\frac{n_x}{N}$ of the frequencies of each array which it crosses. The frequency of the $x$ array of type $y$ is $n_y$. Hence the subgroup of intersection has a frequency, $\frac{n_x}{N} \cdot n_y$, which equals $\frac{n_x \cdot n_y}{N}$.

Now if in the actual distribution the frequency of a subgroup, $n_{xy}$, is larger or smaller than the random selection frequency given by the formula $n_x \cdot \frac{n_y}{N}$, the divergence must be due to the presence in the data of a tendency for *certain* values of the attributes to be *most often* associated and hence the total extent of this divergence is a measure of the degree of the **association** or correlation in the data. This method of measuring correlation is called the **method of contingency.**

The difference $n_{xy} - \frac{n_x \cdot n_y}{N}$ is squared to prevent the cancelling of positive and negative values. Since only the relative size of the differences is significant, this square is divided by the above random selection frequency $\frac{n_x \cdot n_y}{N}$. On summing all such values, we have the **mean square contingency** $\phi$.[2]

where $N\Phi^2 = \Sigma\Sigma \dfrac{\left(n_{xy} - \dfrac{n_x n_y}{N}\right)^2}{\dfrac{n_x n_y}{N}}$

On expanding and reducing, this summation is arranged in a more convenient form for computation. We have

$$\frac{\left(n_{xy} - \dfrac{n_x n_y}{N}\right)^2}{\dfrac{n_x n_y}{N}} = N\frac{n_{xy}^2}{n_x n_y} - 2n_{xy} + \frac{n_x n_y}{N}.$$

and hence $\Sigma\Sigma \dfrac{\left(n_{xy} - \dfrac{n_x n_y}{N}\right)^2}{\dfrac{n_x n_y}{N}} = N\Sigma\dfrac{n^2_{xy}}{n_x n_y} - 2N + N,$

since $\Sigma\Sigma \dfrac{n_x n_y}{N} = \dfrac{1}{N}\Sigma n_x \Sigma n_y = \dfrac{1}{N}\Sigma n_x N = \Sigma n_x = N.$

Therefore $\phi^2 = \Sigma\dfrac{n^2_{xy}}{n_x n_y} - 1.$

The probable deviation of $\phi$ is discussed at length by Pearson and Blakeman*.

## Exercises.

1. Compute the value of $\phi$ for the data of Table VIII.
2. Why not divide the square of the difference for each sub-group by the actual frequency of the sub-group instead of the frequency under the assumption of no correlation?

**Properties of $\phi$.** In data selected entirely at random; that is, where $n_{xy} = \dfrac{n_x n_y}{N}$ for all values of $x$ and $y$, the value of $\phi$ is of course zero. It does not necessarily follow, however, that for absolutely uncorrelated material; that is, for data having $\eta_x = \eta_y = 0$, the value of $\phi$ must be zero.

A moment's consideration will show that the greatest value for $\Sigma\Sigma\dfrac{n^2_{xy}}{n_x n_y}$, taken over the subgroups of any one array, is unity and that this greatest value cannot be attained unless the subgroup of intersection is the only subgroup with non-vanishing frequency in either of the two arrays intersecting in that subgroup. It follows that, if the distribution is not square, the number of arrays giving the maximum value cannot be greater than the number of the longer arrays. Hence in symbols, if $r$ and $s$ are the numbers of arrays of the respective attributes and $r = s$ or $r < s$, the greatest value for $\phi^2$ is, $r - 1$.

* Biometrika. Vol. V, p. 191 et seq.

For illustration, in the table

$$\begin{array}{cccc} a & b & c & d \\ e & f & g & h \\ i & j & k & l \end{array}$$

at least one horizontal array must have more than one non-vanishing subgroup frequency. Let this table be

$$\begin{array}{cccc} a & o & o & o \\ o & f & o & o \\ o & o & k & l\,. \end{array}$$

$$\text{Then } \Sigma \frac{n^2_{xy}}{n_x n_y} = \frac{a^2}{a.a} + \frac{f^2}{f.f} + \frac{k^2}{k(k+l)} + \frac{l^2}{l(k+l)}$$

$$= 1 + 1 + \frac{k}{k+l} + \frac{l}{k+l}$$

$$= 3.$$

## Exercises.

3. Show that for $\phi = 0$, the means of the $x$ and the $y$ arrays lie on vertical and horizontal lines respectively.

4. Show by actual substitution in the formula $\phi^2 = \Sigma \frac{n_{xy}^2}{n_x n_y} - 1$ that $\phi = 0$ when $n_{xy} = \frac{n_x n_y}{N}$.

5. Verify the just preceding theory by assigning different combinations of values to the symbols $a, b, c, d, e, f, g, h, i$, in the distribution:

$$\begin{array}{ccc} a & b & c \\ d & e & f \\ g & h & i \end{array}$$

6. Do the same for the distribution

$$\begin{array}{ccccc} a & b & c & d & e \\ f & g & h & i & j\,. \end{array}$$

7. Show that the greatest value of $\phi^2$ for a table of the form of Table VIII is 13.

8. Give an algebraic demonstration for this theory when applied to a general distribution $r$ by $s$ fold.

The **greatest disadvantage** of $\phi$ as a measure of correlation arises from the fact that its value depends on the number of arrays in the distribution so that it is almost entirely useless for purposes of comparison Another disadvantage lies in the fact that, notwithstanding the logical simplicity and directness of the theory underlying the method of contingency, in practice the interpretation of variations in the value of $\phi$ is a matter of much difficulty For instance, when $\phi$ equals 2.5 what is the significance of an increase of 0.5 in its value? How much greater is the degree of closeness of association in the latter case than in the first? A third objection is that for a large table the labor of computation is heavy.

The first objection above is partially overcome by making use of the *coefficient of contingency*, $\sqrt{\frac{\phi^2}{1+\phi^2}}$. This constant is given added prestige by the following relation. It may be shown that for a *finely divided distribution of a particular type* the *coefficient of contingency* and the *coefficient of correlation are equal in value.* Consequently in certain forms of distribution this furnishes a convenient method of obtaining the value of $r$. However, care must be taken to make sure that the assumptions essential for the validity of this theorem are approximated to with sufficient closeness. *Ordinarily it is better to make use of methods which do not rest on so extensive assumptions.*

An approximation to the **probable deviation** of the coefficient of contingency is to take one and one-third the probable deviation of r.

## Exercises.

9. Compute the coefficient of contingency for the data of Table VIII and compare with the value of $r$ already computed.

10. Do the same for the data of Table IX, Chapter VII.

11. By combining arrays in the distribution of Table VII and computing the successive values of $\sqrt{\frac{1+\phi^2}{\phi^2}}$ show the effect of different widths of classes on the value of this constant.

12. Show that the coefficient of contingency is smaller than the value of $r$ computed by the method of moments.

13. Compare the reliability of the coefficient of contingency for highly and for slightly correlated data.

14. Compare the labor required to compute the value of $\phi$ with that for $\eta$.

In concluding this part of the discussion of the method of contingency it may be stated that *when the attributes can be definitely measured there is no practical advantage in computing the value of* $\phi$.

**Non-Quantitative Characteristics.** Because the formula $\Sigma\Sigma \frac{n^2_{xy}}{n_x n_y}$ does not contain the deviations $x$ and $y$ and contains only the frequencies of the subgroups it can be applied to distributions in which it is impossible or undesirable to assign numerical values to the deviations; for instance, a distribution of hair and eye color, of degrees of intelligence in drawing and music. Such distributions are said to involve characteristics not *quantitatively measured or measurable.*

Thus in its fundamental theory the coefficient of contingency applies with equal validity to quantitative and to non-quantitative data. Moreover, since the number of classes in the case of non-quantitative distributions is ordinarily small the labor of computation is not unduly heavy, and hence the coefficient of contingency is of greater practical importance for this kind of data than for quantitative data. However it will now be shown that for non-quantitative distributions, the correlation ratio is a more convenient and satisfactory measure of correlation than is the coefficient of contingency.

A correlation problem very similar to that arising from non-quantitative data is the finding of the degree of correlation when the measurements of the attributes in quantitative data are classified into very broad classes; to find, for instance, the extent of the tendency for under-height and over-weight to be associated. Further than the effect that so broad classes may have in producing errors in the results obtained by the formula for $\eta$ there is no *theoretical objection to the direct application of the theory of the correlation ratio to a distribution obtained by grouping into broad classes.*

However, *the theory of the correlation ratio does not apply directly to strictly qualitative data and for that reason we shall justify its use for such distributions by showing that*

*in a very important form of distribution, the two by two table, $\eta$ and $\phi$ are identical and that in other ordinarily occuring cases the values of the two constants are highly correlated.*

## Exercises.

16. Arrange the data of Table VIII in the following form and compute the values of $\eta$ and $\phi$.

Height.

| Weight. | | Under 68 | Over 68 | Totals |
|---|---|---|---|---|
| | Under 137 | | | |
| | Over 137 | | | |
| | Totals | | | |
| | | | | |

**The Four-fold and the Nine-fold tables.** We shall now derive the formulas for $\eta$ and $\phi$ for a $2 \times 2$ table and obtain the computation formulas for $\eta$ for a 3 by 3 table. The same method might be employed to derive special formulas for each type of table. In the absence of special formulas the general formula for $\eta$ can be applied directly.

Let us take the four-fold table,

| | |
|---|---|
| $n_{11}$ | $n_{21}$ |
| $n_{12}$ | $n_{22}$ |

We have $\bar{y} = \dfrac{n_{:1} + 2n_{:2}}{N} = \dfrac{n_{:2}}{N} + \dfrac{n_{:1} + n_{:2}}{N} = 1 + \dfrac{n_{:2}}{N}$,

Similarly, $\bar{y}_1 = 1 + \dfrac{n_{12}}{n_1}$, and $\bar{y}_2 = 1 + \dfrac{n_{22}}{n_2}$.

Substituting these values in the formula,

$N\sigma m_y^2 = n_1(\bar{y}_1 - \bar{y})^2 + n_2(\bar{y}_2 - \bar{y})^2$, where $\sigma m_y^2$ is the mean squared deviation of the means of the arrays,

We have after some detailed reduction,

$$N\sigma m_y^2 = \frac{(n_{11}n_{22} - n_{12}n_{21})^2}{Nn_1 . n_2}.$$

Also $N\sigma_y^2 = \frac{n_{:1}n_{:2}}{N}$.

Therefore $\eta = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_1 n_2 n_{:1} n_{:2}}}$.

From the formula $N\sigma^2 = \Sigma\Sigma \frac{n_{xy}^2}{n_x n} - 1$ it is readily shown by direct computation that $\phi = \frac{m_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_1 n_2 n_{:2} m_{:1}}}$.

*The equality of $\phi$ and $\eta$ for the fourfold distribution* is therefore demonstrated.

For the nine-fold table,

| | | |
|---|---|---|
| $n_{11}$ | $n_{21}$ | $n_{31}$ |
| $n_{12}$ | $n_{22}$ | $n_{32}$ |
| $n_{13}$ | $n_{23}$ | $n_{33}$ |

we have by a reduction similar to that for the 2 by 2 table,

$$\bar{y} = 1 + \frac{n_{:2} + 2n_{:3}}{N}, \bar{y}_1 = 1 + \frac{n_{12} + 2n_{13}}{N_1}, \bar{y}_2 = 1 + \frac{n_{22} + 2n_{23}}{N_2},$$

$$\bar{y}_3 = 1 + \frac{n_{32} + 2n_{33}}{N_3}.$$

Let $\frac{n_{:2} + 2n_{:3}}{N} = l,$

then, on substituting for $\bar{y}_x$ and $\bar{y}$,

$$n_x(\bar{y}_x - \bar{y})^2 = n_x \left( \frac{n_{x2} + 2n_{x3}}{n_x} - l \right)^2,$$

$$= \frac{(n_{x2} + 2n_{x3})^2}{n_x} - 2l \left( \frac{n_{x2} + 2n_{x3}}{N} \right) N + n_x l^2.$$

Therefore, $\Sigma n_x(\bar{y}_x - \bar{y})^2 = \Sigma \frac{(n_{x_2} + 2n_{x_3})^2}{n_x} - 2lN\Sigma \frac{2n_{x_2} + 2n_{x_3}}{N}$

$+\Sigma n_x l^2 = \Sigma \frac{(n_{x_2} + 2n_{x_3})^2}{n_x} - 2Nl_2^2 + Nl^2, = \Sigma \frac{(n_{x_2} + 2n_{x_3})^2}{n_x} - Nl^2.$

Similarly $\Sigma n_y(y - \bar{y})^2 = N(l - l^2) + 2n_{:1}.$

Hence $\eta^2 = \frac{\Sigma \frac{(n_{x_2} + 2n_{x_3})^2}{n_x} - Nl^2}{N(l - l^2) + 2n_{:3}}$

## Exercises.

17. Compute the value of $\eta$ from the following distribution of the variation in receipts and prices from month to month of live hogs at Union Stock Yards, Chicago, from 1901 to 1914.

Receipts.

| Prices. | ..... — 50 | —50....50 | 50..... |
|---|---|---|---|
| ..... — 25 | 10 | 7 | 24 |
| 25.....25 | 20 | 32 | 26 |
| 25..... | 37 | 5 | 6 |

The original formula for $\phi$,

$$\phi^2 = \Sigma\Sigma \frac{n^2_{xy}}{n_x n_y} - 1$$

is probably in the case of the 3 x 3 table as convenient as any for the computation of that constant.

It is immediately evident that in general $\eta$ *and* $\phi$ *cannot be equivalent* for a table larger than four-fold, because there are two $\eta$'s for each distribution and only one $\phi$. The following theorems may be stated.

1. *When* $\phi = 0$, *the value of* $\eta$ *for* $y$ *on* $x$ *and for* $x$ *on* $y$ *are both zero; that is,* $\eta_y = \eta_x = 0$.

2. *When* $\eta_y = \eta_x = 0$, *it may ordinarily be expected that* $\phi$ *will be practically zero but it is not absolutely necessary that such be the case.*

3. *When only one $\eta$ is zero, it is most likely that $\phi$ will be small in value.*

4. *When $\phi$ takes the maximum value, $\eta_y = \eta_x = 1$.*

5. *When $\eta_y = \eta_x = 1$, $\phi$ takes the maximum value.*

6. *When one $\eta$ only is unity it is most likely that the value of $\phi$ will not differ greatly from the maximum.*

7. *There is a close correspondence between the values of $\phi$ and the $\eta$'s for data of all degrees of correlation.*

**Discussion of the Theorems.** On substituting the relations $n_{xy} = \frac{n_x n_y}{N}$ in the formula for $\eta_y$ and for $\eta_x$ it follows immediately that $\bar{y} = \bar{y}_1 = \bar{y}_2 = \bar{y}$ and hence that $\eta_y = \eta_x = 0$ for $\phi = 0$.

In regard to theorem 2, it will now be shown that the nine relations, $n_{xy} = \frac{n_x n_y}{N}$, which result for a nine-fold table when $\phi = 0$, can be reduced to four independent relations, which result when $\phi = 0$. That is, if there are four such relations the other five must hold true and the value of $\phi$ is necessarily zero. In other words, *the vanishing of $\phi$ imposes four and only four restrictions or conditions on the data of a 3 by 3 table.*

For if, $n_{11} = \frac{n_1 n_{:1}}{N}$, $n_{21} = \frac{n_2 n_{:1}}{N}$, $n_{12} = \frac{n_1 \cdot n_{:2}}{N}$, and $n_{22} = \frac{n_2 n_{:2}}{N}$ it follows that $n_{31} = \frac{n_3 \cdot n_{:1}}{N}$. Let us substitute the equivalents for $n_{11}$ and $n_{21}$ in the equation $n_{31} = n_{:1} - n_{11} - n_{21}$. This substitution gives

$$n_{31} = n_{:1} - \frac{n_1 \cdot n_{:1}}{N} - \frac{n_2 \cdot n_{:1}}{N} = n_{:1} - \frac{n_{:1}}{N}(n_1 + n_2),$$

$$= n_{:1} - \frac{n_{:1}}{N}(N - n_3) = n_{:1} - n_{:1} + \frac{n_3 \cdot n_{:1}}{N},$$

$$= \frac{n_3 \cdot n_{:1}}{N},$$

and similarly for the remaining relations.

The vanishing of $\eta_y$ implies the three relations,

$$\frac{n_{12}+2n_{13}}{n_1}=\frac{n_{22}+2n_{23}}{n_2}=\frac{n_{32}+2n_{33}}{n_3}=\frac{n_{:2}+2n_{:3}}{N}.$$

It can be readily shown that only two of three relations are independent. That is, if the first two relations hold, the third is necessarily true.

If $\eta_x$ as well as $\eta_y$ vanish the three additional relations $\frac{n_{21}+2n_{31}}{n_{:1}}=\frac{n_{22}+2n_{32}}{n_{:2}}=\frac{n_{32}+2n_{33}}{n_{:3}}=\frac{n_2+2n_3}{N}$ are implied. Here again only two of the three additional relations are independent and of the six relations implied by the vanishing of both $\eta_y$ and $\eta_x$ it is only a matter of algebraic detail to show that only three are independent. That is, the vanishing of both $\eta_x$ and $\eta_y$ imposes one less condition on the data than does the vanishing of $\phi$. And hence it is not necessarily true that $\phi = 0$ when $\eta_x = \eta_y = 0$.

As to the maximum values for these constants, the relations $\eta_y = \eta_z = 1$ require that there be but one non-vanishing frequency in each array of either sense and hence the condition for a maximum value for $\phi$ is satisfied. The converse relations are evidently true.

For only one $\eta$ equal to unity, however, the data might be arranged, for instance, in the form of the table,

| | | |
|---|---|---|
| a | o | o |
| o | o | o |
| o | b | c, |

when $\phi^2$ would not have the maximum value.

If a large number of distributions were made up from the same population and the values of $\eta$ and of $\phi$ computed for each distribution, it would be found that in the long run a large value of $\eta$ was associated with the larger values of $\phi$ and vice versa. But to obtain a formula for the correlation of $\eta$ and $\phi$ is a matter of considerable algebraiac detail and the resulting formula is so complicated that it is practically worthless*. For this reason the algebraiac discussion of Theorems 2, 3 and 6 is not given in the complete form.

We have outlined the method of showing that *the value of $\eta$ for a non-quantitative distribution has a close connection to the*

* Compare Blakeman, "The Probable Error of the Coefficient of Contingency" loc. cit.

*value of $\phi$; that is, that $\eta$ and $\phi$ are highly correlated, for such data and hence the correlation ratio may be used to measure the degree of correlation or association in the data with all the assurance that attaches to the method of contingency.* It is of distinct practical advantage to have one coefficient or index of correlation for all kinds of data and for that reason the coefficient of contingency is not greatly used in practice.

Caution is necessary at one point, however, for *data divided into only a few classes does not convey the same amount of information regarding the correlation between the characteristics as does the more detailed material and hence not the same degree of confidnce can be placed in the computed value of any constant derived from the less detailed table. For this reason comparisons of the values of correlation measures between different forms of distributions must be carefully made and due account taken of the fact that for the small table the results do not warrant the same degree of confidence as do the results from the finely divided table.*

## Exercises.

18. Prove that if

$$n_{11} = \frac{n_{1.}n_{:1}}{N},$$

$$n_{21} = \frac{n_{2.}n_{:1}}{N},$$

$$n_{12} = \frac{n_{1.}n_{:2}}{N},$$

and $$n_{22} = \frac{n_{2.}n_{:2}}{N},$$

the remaining five relations of the same type hold.

19. Show that if $\eta_2 = 0$ the vanishing of $\eta_x$ imposes only one additional condition on the data.

20 Show that if $\eta_y = \eta_x = 0$, the frequencies of the nine fold table, can be expressed in terms of the marginal sums and frequency of any one sub-group.

21. Show that in the distribution $\begin{matrix} 2 & 4 & 2 \\ 1 & 1 & 1 \\ 2 & 4 & 2 \end{matrix}$ $\eta_y = \eta_x = 0$ and $\phi \pm 0$.

22. Construct a fictitious table having $\eta_y = 1$ and $\phi$ not having a maximum value.

23. Investigate the relations between $\eta$ and $\phi$ for a $2 \times 3$ table.

## APPENDIX I.

**Introduction.** The generalized frequency curves of Pearson are so diverse in shape that a curve of this class can be found to fit any ordinary statistical distribution. By the following methods the fitting of a Pearson curve is reduced almost entirely to a matter of routine substitution in formulas, so that the practical statistician can make extended use of the curves without great familiarity with their theory.

This discussion is designed both to present the working methods of the generalized frequency curves and to give the statistician who has a minimum of acquaintance with the higher mathematics some degree of familiarity with the underlying theory. The demonstrations are, for the most part, omitted. Many of the exercises have to do with the omitted theorems and derivations.

In developing the theory of the generalized frequency curves it is logical, as well as practically convenient, to start with the normal curve and consider the *general distribution as a modification* of the normal type of distribution.*

**The Slope Property.** The particular modification which leads to the frequency of Pearson is obtained by generalizing the slope condition of the normal curve.** The *slope* of a *curve* at a given point is the tangent of the angle which the line touching the curve at that point makes with the X-axis. In the case of the normal curve, the ratio of the slope to the ordinate is negatively equal to the abscissa of the point.

This slope property is generalized by taking the ratio equal, not to $-x$, but to $-(x+a)(b+cx+dx^2)$ where $a$, $b$, are equal, not to $-x$, but to $-\frac{x+a}{b+cx+dx^2}$ where $a$, $b$, $c$, $d$, are constants. The slope of a curve is ordinarily denoted by the symbol $\frac{dy}{dx}$.

* Compare Edgeworth, Jour. Roy. Stat. Soc. Also West, "On the Translated Normal Curve," Ohio Journal of Science, Dec., 1915.

** First extensively treated by Pearson in the article "Skew Variation in Momogeneous Material," Phil. Trans.

In this notation the generalized slope property is expressed by the equation.

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x+a}{b+cx+dx^2}.$$

**The Constants, a, b, c, d.** The statistical significance of each of the constants, $a$, $b$, $c$, $d$, can be readily determined.

In Chapter IV, it is shown that the slope of a frequency curve is zero at a mode. Since $\frac{dy}{dx}$; that is, the slope, is zero when $x = -a$, the constant $a$ determines the position of the mode. The mode is therefore at a distance, $-a$, from the mean. As explained in Chapter V. $a$ is thus a measure of the skewness, of the lack of symmetry of the distribution. For a symmetrical distribution $a$ is evidently $o$.

When both $c$ and $d$ are zero the generalized slope equation is merely the normal slope equation with $x$ replaced by $\frac{x+a}{b}$. This leads to the normal curve,

$$y = k \cdot e^{-\frac{(x+a)^2}{b}}, \text{ where } k \text{ is a constant.}$$

Comparing this equation with the standard normal equation,

$$y = k\, e^{-\frac{x^2}{2\sigma^2}}.$$

we see that $b$ equals $2\sigma^2$ multiplied by a constant.

The degree of symmetry of the curve is indicated by the value of $c$ as well as by the value of $a$. For, when $x$ is positive, the term $cx$ is added in the denominator and when $x$ is negative it is subtracted. This tends to make the frequency curve steeper to the left than to the right of the origin, and hence the curve must extend farther to the right, that is, the curve must be skew.*

But it was seen in Chapter V. that $\beta_1$ is the fundamental measures of skewness. Therefore both $a$ and $c$ must contain $\beta_1$ as a factor.

When $x^2$ is small the constant $d$ has little effect on the

* See page 57, Chapter V.

slope, but for the extremities of the curve where $x$ and hence $dx^2$ is large the slope is reduced by a large value of $d$. It will be seen that $d$ depends largely on $\beta_2$.

**The Types of Curves.** We may now discuss the distinct types of curves that possess the slope properties of the generalized slope equation. *Distinct types of curves result according as the denominator, $b + cx + dx^2$, has two distinct factors, two coincident factors, or has no factors.*

With two distinct factors the slope equation can be written

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x+a}{b+cx+dx^2} = k \cdot \frac{x+a}{(r+x)(r^2-x)}$$

where $k$ is a constant.

By the usual mathematical methods we then have

$$y = y' (r_1 + x)^{\frac{k(a-r_1)}{r_1+r_2}} \cdot (r_2 - x)^{\frac{-k(a+r_2)}{r_1+r_2}} \quad (A)$$

where $y'$ is the constant of integration.

By a simple transformation and rearrangement, this equation can be reduced to the form of Pearson's first type, namely:

$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2}. \qquad \text{Type I.}$$

### Exercises.

1. Carry through in detail the necessary transformations to determine the equation of Type I from equation (A).
2. Perform the integrations to obtain the curve of Type I.

When $a_1$ and $a_2$ are equal it is readily shown that $m_1 = m_2$ and the equation takes the form of Type II:

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^m. \qquad \text{Type II.}$$

When one root of the denominator $b + cx + dx^2$ is indefinitely large, that is, when $d$ is zero, we have, from the theory of the exponential e, the third type:

$$y = y_0 e^{-\gamma x} \left(1 + \frac{x}{a}\right)^{\gamma a}. \qquad \text{Type III.}$$

This equation may be looked upon as that of Type I with $a_2$ indefinitely large.

The curves of Type III are especially serviceable because the equations are simple in form and convenient for computation. They are the most elementary skew curves.

By transforming expression $(A)$, in a manner somewhat different from that to obtain Type I, the form of Pearson's sixth type is readily obtained. It is

$$y = y_0 (x - a)^{m_2} x^{-m_1}. \qquad \text{Type VI.}$$

### Exercises.

3. Obtain the equation of Type II by direct integration from the differential equation.
4. Compare Type II with the normal curve.
5. Obtain Type III directly by integration.
6. Obtain Type III from (A).
7. Compare the shape of Type III with that of the normal curve.
8. Obtain the equation of Type VI directly from the differential equation.
10. Is Type VI geometrically distinct from Type I?

When **two roots are indefinitely large** we have the normal curve:

$$y = y_0 \, e^{-\frac{x^2}{2b}}$$

which is called simply "Normal" in Pearson's scheme of classification.

*With two coincident roots,* the slope equation becomes

$$\frac{1}{y} \cdot \frac{dy}{dx} = k \frac{x + a}{(x + r)^2}$$

This leads to the form $y = y_0 x^{-p} e^{-\frac{\gamma}{x}}$, Type V.
which is Pearson's type V.

### Exercises.

11. Derive in detail the equation of Type V.

When the denominator of the slope equation cannot be factored the integration is performed by writing

$$\frac{1}{y}\cdot\frac{dy}{dx}=-\frac{x+a}{b+cx+dx^2},$$

$$=-\frac{x+\frac{c}{2d}+a-\frac{c}{2d}}{d\left(x^2+\frac{c}{d}x+\frac{c^2}{4d^2}+\frac{b}{d}-\frac{c^2}{4d^2}\right)}.$$

This gives

$$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}e^{-\nu\tan^{-1}\frac{x}{a}},$$

which is the form of Type IV. Type IV.

**Exercises.**

12. Derive in detail the equation of Type IV.
13. Derive the equation of Type IV by transformation from the equation of Type I.
14. Compare the form of the equation of Type IV to that of Type III.

If $\gamma$ is zero in the immediately preceding equation we have Pearson's Type VII.

$$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}$$ Type VII.

**The Intercepts.** The intercepts made on the X-axis by the various types of curves can now be examined. The following theorem is fundamental in the theory of the intercepts of Pearson's curves: *an incommensurable power of a negative number does not exist.*

Let $-N$ denote any negative number and $(-N)^p=r(\cos p\pi+\sqrt{-1}\sin p\pi)$ where $\sqrt{-1}$ is the square root of negative unity. Unless $p$ is an integer $\sin p\pi$ is not zero and hence $(-N)^p$ contains $\sqrt{-1}$ which has no arithmetical value. Hence powers of $-N$ which are not integral do not exist.

In Type I the intercepts are $-a_1$ and $a_2$. Since $a_1$ and $a_2$ are not integers, the curve stops at the X-axis and *there are no points below that axis.* Indeed, there are no negative ordinates on any of the curves.

In Type II the intercepts are of the same length and numerically equal to $a$.

In Type III one intercept is $-a$ and the other is indefinitely large.

In the case of the normal curve both intercepts are indefinitely large.

In Types IV and VII there are no intercepts.

In Type V one intercept passes through the origin and the other is indefinitely large.

In Type VI both intercepts are positive or both are negative.

Ordinarily the type of curve selected should have intercepts harmonizing with the natural limits of the range of the data. For instance, data necessarily limited in either direction should be smoothed with a curve correspondingly limited. However nearly all the curves are practically limited in range because the ordinates soon become negligible, so that the matter is not one of great importance; tho a somewhat better fit is likely to be obtained with a curve limited in accordance with the data.

### Exercises.

15. Of what types is the normal curve a limiting curve?

16. Distinguish between a curve with indefinitely large intercepts and a curve with imaginary or non-existent intercepts.

17. Show that there are indefinitely more curves of Types I, VI and IV than of Types III, V, II or VII, or of the normal curve.

18. Show how Type I can be said algebraically to include Type IV.

19. Show that Types I and VI are not fundamentally distinct.

20. Show that by taking all combinations of sign into account there are three distinct classes of curve under Type I.

21. Show that there are two sub-classes under Type II according as the exponent $m$ is positive or negative.

22. Show that there are two classes under Type III.

23. Is there more than one general form of curve under Type IV? Under type V?

24. Discuss the curves of Type VI as to the existence of sub-classes within the Type.

25. What types of these curves have asymptotes?

26. Do all the curves have a mode?

27. Find the points of inflexion for each type.

**The Criterion K.** Since the separation into types depends primarily on the nature of the roots of the quadratic, $b + cx + dx^2$, the discriminant of this quadratic constitutes a

criterion of the type of curve which fits the distribution. The values of $a$, $b$, $c$, and $d$ are first determined by the method of moments and then the discriminant expressed in terms of the computed expressions for $b$, $c$, and $d$.

The formula for K, the discriminant obtained in this way is

$$K = \frac{\beta_1(\beta_2 + 3)^2}{4(2\beta_2 - 3\beta_1 - 6)(4\beta_2 - 3\beta_1)}$$

This formula for K is derived as follows:

The differential equation $1/y\ dy/dx = -(x + a)/(b + cx + dx^2)$ may be written $(b + cx + dx^2)\ dy/dx = y(x + a)$. Multiplying each side by $x^n$, we have

$\int x^n(b + cx + dx^2)dy = -\int y(x + a)x^n dx$. On integrating the left side by parts

$x^n(b + cx + dx^2)y - nb\int x^{n-1} y\ dx - (n+1)c\int x^n y\ dx - (n+2)d\int x^{n+1} y\ dx = -\int y\ x^{n+1}\ dx - a\int y\ x^n dx$.

With the usual notation, where $\mu'_n = \int x^n y\ dx$,

$x^n(b + cx + dx^2)y - nb\mu'_{n-1} - (n+1)c\mu'_n - (n+2)d\mu'_{n+1} = -\mu'_{n+1} - a\mu'_n$.

If $y$ is very small at the ends of the range the first expression vanishes and the moment equation connects the three moments $\mu'_{n-1}$, $\mu'_n$, and $\mu'_{n+1}$.

On rearranging this equation we have

$a\mu'_n - nb\mu'_{n-1} - (n+1)c - (n+2)d\mu'_{n+1} = -\mu'_{n+2}$.

Since the moment, $\mu'_0 = 1$ and, if the mean is taken as origin, $\mu'_1 = 0$ we have for $n = 0, 1, 2, 3$, respectively the four equations:

$$\begin{aligned} a - c &= 0 \\ b - 3d\mu_2 &= -\mu_2 \\ a\mu_2 - 3c\mu_2 - 4d\mu_3 &= -\mu_3 \\ a\mu_3 - 3b\mu_2 - 4c\mu_3 - 5d\mu_4 &= -\mu_4 \end{aligned}$$

On solving this set of equations and substituting in the differential or slope equation, we have

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x + \dfrac{\mu_3(\mu_4 + 3\mu_2^2)}{10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2}}{\dfrac{\mu_2(4\mu_2\mu_4 - 3\mu_3^2)}{10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2} + \dfrac{\mu_3(\mu_4 + 3\mu_2^2)}{10\mu_3\mu_4 - 18\mu_2^3 - 12\mu_3^2}x + \dfrac{2\mu_2\mu_4 - 3\mu_3^2}{10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2}x^2}.$$

In terms of $\beta_1$ and $\beta_2$ this becomes

$$\frac{1}{y}\frac{dy}{dx} = -\frac{x + \dfrac{\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}}{\dfrac{\mu_2(4\beta_2 - 3\beta_1) + \sqrt{\mu_2}\sqrt{\beta_1}(\beta_2 + 3)x + (2\beta_2 - 3\beta_1 - 6)x^2}{2(5\beta_2 - 6\beta_1 - 9)}}$$

The discriminant of the quadratic denominator is the required criterion, K. It is easily shown that

$$K = \frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)}.$$

For, the quadratic expression, $dx^2 + cx + b$, may be written $d\left(x - \frac{\sqrt{c^2-4bd}}{2d}\right)\cdot\left(x + \frac{\sqrt{c^2-4bd}}{2d}\right)$. Hence the character of the two factors depends on the value of the quantity $(c^2-4bd)$. When this is zero the two factors are equal; when it is negative there are no factors, etc. Writing $(c^2-4bd)$ in the form $(c^2/4bd-1)\,4bd$ we have, if $K = \frac{c^2}{4bd}$, the following classes of factors according to values of K:

If $K < 0$; that is, if K is negative, the factors are unequal, because a negative sign for $c^2/4bd$ must be due to unlike signs for $b$ and $d$ and hence the product $4bd$ must be negative. That is, $c^2/4bd$ is positive for $K < 0$.

If K is positive there are two cases, according as K is greater or less than unity. If K lies between zero and positive unity, $(c^2-4bd)$ is negative and consequently there are no factors. If $K > 1$, $(c^2-4bd)$ is again positive and the factors are unequal, etc.

**The Value of K and the Types of Curve.** The following table gives the types of curves corresponding to the different values of $K$.

| | | |
|---|---|---|
| $K < 0$, | i. e. negative | Type I. |
| $K = 0$, | $\beta_1 = 0,\ \beta_2 = 3$ | Normal Curve. |
| | $\beta_1 = 0,\ \beta_2 < 3$ | Type II. |
| | $\beta_1 = 0,\ \beta_2 < 3$ | Type II. |
| $K > 0 < 1$ | | Type IV. |
| $K = 1$ | | Type V. |
| $K >$ | 1, but not indefinitely large, | Type VII. |
| $K >$ | 1 and indefinitely large. | Type III. |

It is to noted that the types of curve for any given statistical distribution can now be determined by strictly arithmetic methods.

The only restriction on the generality of the theory of the criterion K is that the quantity $x^n(b + cx + dx^2)y$ must vanish at both ends of the range. This condition marks the pairs of values of $\beta_1$ and $\beta_2$ for which no curve of the generalized differential equation can be found. The limiting values of $\beta_1$ and $\beta_2$ are $\beta_2 > \frac{3}{4}\beta_1$ and $\beta_2 > \beta_1/8 + 9/2$ (see Exercises 29 and 30 below).

## Exercises.

28. Read the explanation to Tables XXXV-XLVI in "Tables for Statisticians and Biometricians."

29.* Derive the formulas

$$\beta_n(\text{even}) = (n+1)\ \{1/2\beta_{n-1} + (1+\tfrac{1}{2}a)\beta_{n-2}\} / \{1-\tfrac{1}{2}(n-1)a\}$$
$$\beta_n(\text{odd}) = (n+1)\ \{1/2\beta_1\beta_{n-1} + (1+\tfrac{1}{2}a)\beta_{n-2}\} / \{1-\tfrac{1}{2}(n-1)a\}$$
where $a = (2\beta_2 - 3\beta_1 - 6) / (\beta_2 + 3)$.

30. From the computation formulas for Type II, prove that $m$ is negative when $\beta_2 < 1.8$.

31. Prove from the working formulas of Type I that Type I includes three sub-classes according to the signs of $m_1$ and $m_2$. Derive the criterion curve,

$$\beta_1(8\beta_2 - 9\beta_1 - 12)\ (4\beta_2 - 3\beta_1) = (10\beta_2 - 12\beta_1 - 18)^2\ (\beta 2 + 3)^{2*}$$

32. Prove that $\beta_2 > \beta_1$

33. Prove the relation $\beta_2 > 15/8\beta + 9/2$.

34. Show that a large value of $\beta_2$ for the curves derived from the generalized differential equation denotes a comparatively flat-topped curve.

35. Show that for the normal curve with $a = o$, we have $b = \sigma^2$, $c = d = o$, $\mu_4 = 3\sigma^4$.

**The Computation Formulas.** The computation formulas for the several types of Pearson's frequency curves are derived in accordance with the method of moments. For each type as many moment equations are written as there are constants in the equation of a curve of the type. In some of the type equations, as in Type I where $a_1/m_1 = a_2/m_2$, the constants are connected by equations so that the number of moment equations is reduced. *The moment equations are the result of equating the theoretical moments of the curve obtained by integration to the moments computed directly from the data.*

It might be expected that the differential equation in terms of $\mu_2$ and the $\beta$'s would be integrated to give the equations directly, but the present process is more convenient. The chief purpose, therefore, of the slope or differential equation is for the determination of the type forms of the equations. After the algebraic forms of the equations are determined each type is worked out without making use of its connection either with the slope equation or with other type forms.

---

* See page lxiii, of "Tables for Statisticians and Biometricians."

The expression $\Gamma(p)$, called the *gamma function,* occurs in the following formulas. This function is defined by the relation

$\Gamma(p) = (p-1)\,\Gamma(p-1)$.

If $p$ is an integer, $\Gamma(p) = \underline{|p-1}$.

If $p$ is not an integer, $\Gamma(p) = (p-1)(p-2) \ldots (p-p+2)\,\Gamma P$ where P is the remainder after subtracting a sufficient number of 1's to bring $p$ down to between 2 and 1 in value. The values of $\Gamma(P)$ are given in Table XXXI of "Tables."

The probable errors of K as well as of $\beta_1$ and $\beta_2$ are given in "Tables."

The derivation of the following computation formulas, except the moment formulas, is not possible without an extensive acquaintance with the calculus.*

After the constants in the equation are computed the smoothed frequencies are obtained by computing the areas under the curve and between the bounding ordinates. Thus the frequency of the first class is the area between the ordinate $x = \frac{1}{2}$ and $x = 1\frac{1}{2}$. Simpson's quadrature formula is ordinarily used for finding the class areas. According to this formula the area is $1/6\,\{y_{x-\frac{1}{2}} + 4y_x + y_{x+\frac{1}{2}}\}$ where $y_{x-\frac{1}{2}}$ and $y_{x+\frac{1}{2}}$ are the bounding ordinates and $y_x$ is the mid-ordinate of the class.

**Formulas for the Moments.**

$S_2 = d$.

$\nu_2 = 2S_3 - d(1+d)$.

$\nu_3 = 6S_4 - 3\nu_2(1+d) - d(1+d)(2+d)$.

$\nu_4 = 24S_5 - 2\nu_3\{2(1+d)+1\} - \nu_2\{6(1+d)(2+d) - 1\} - d(1+d)(2+d)(3+d)$.

$\mu_2 = \nu_2 - \frac{1}{12}$

$\mu_3 = \nu_3$

$\mu_4 = \nu_4 - \frac{1}{2}\nu_2 + \frac{7}{240}$

$\sigma = \sqrt{\mu_2}$

$\beta_1 = \mu_3^2 \div \mu_2^3$

$\beta_2 = \mu_4 \div \mu_2^2$

$$K = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}$$

* See Elderton "Frequency Curves and Correlation," C. & E. Layton, for a thoro discussion of the deriviations.

**The computation formulas for Type I are as follows:** The equation is,

$$y = y_0 \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2}$$

where $a_1/m_1 = a_2/m_2$.

We have

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{3\beta_1 - \beta_2 + 6}$$

$$\epsilon = \frac{4r^2}{16(r+1) + \beta_1(r+2)^2}$$

$$b^2 = \frac{\mu_2(r+1)r^2}{\epsilon}.$$

$m_2$ and $m_1$ are given by the formulas

$$\tfrac{1}{2}(r-2) \pm \tfrac{1}{2}(r+2)\sqrt{\beta_1\epsilon}.$$

The constant $m_1$ is taken with the negative root when $\mu_3$ is positive and with the positive root when $\mu_3$ is negative.

$a_1 + a_2 = b$.

$a_1$ and $a_2$ can be found from the relations $a_1 + a_2 = b$ and $a_1/m_1 = a_2/m_2$.

$$y_0 = \frac{N}{b} \frac{m_1^{m_1} m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}} \frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}$$

$$\text{The skewness is } \tfrac{1}{2}\sqrt{\beta_1}\left\{\frac{r+2}{r-2}\right\}$$

$$\text{Mode} = \text{mean} - \tfrac{1}{2}\frac{\mu_3}{\mu_2}\left\{\frac{r+2}{r-2}\right\}$$

**The formulas for Type II are as follows** The equation for this type is

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^m$$

The formulas are

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)}$$

$$a^2 = \frac{2\mu_2\beta_2}{3 - \beta_2}$$

$$y_0 = \frac{N\,\Gamma(2m + 2)}{a.2^{2m+1}\{\Gamma(m + 1)\}^2}$$

**Type III.** The equation is

$$y = y_0\, e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a}.$$

The formulas are,

$$\gamma = 2\frac{\mu_2}{\mu_3},$$

$$a = \mu\gamma - \frac{1}{\nu},$$

$$y_0 = \frac{N}{a}\frac{p^{p+1}}{e^p\Gamma(p + 1)}, \quad \text{where } p = \nu a.$$

$$\text{Mode} = \text{mean} - \frac{1}{\gamma}$$

$$\text{Skewness} = \frac{1}{\sigma\gamma}$$

**Type IV.** The equation is

$$y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m} e^{-\gamma \tan^{-1}\frac{x}{a}},$$

The formulas are:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{2\beta_2 - 3\beta_1 - 6}$$

$$m = \tfrac{1}{2}(r + 2),$$

$$e = \frac{4r^2}{16(r - 1) - \beta_1(r - 2)^2},$$

$$\nu = \tfrac{1}{2}(r - 2)\sqrt{\beta_1 e},$$

$$a = \frac{1}{2}\frac{r\sigma}{\sqrt{e}},$$

$$y_0 = \frac{N}{a}\sqrt{\frac{r}{2\pi}}\,\frac{e^{\frac{\cos^2\phi}{3r} - \frac{1}{12r} - \nu\phi}}{(\cos\phi)^{r+1}}, \text{ where } \tan\phi = \frac{\nu}{r}.$$

$$\text{Origin} = \text{mean} + \frac{\nu a}{r}.$$

$$\text{Mode} = \text{mean} - \frac{1}{2}\frac{\mu_3(r-2)}{\mu_2(r+2)}.$$

**Type V.** The equation is

$$y = y_0 x^{-p} e^{-\gamma/x}$$

The formulas are:

$$p = 4 + \frac{8 + 4\sqrt{(4+\beta_1)}}{\beta_1},$$

$\gamma = (p-2)\sqrt{\mu_2(p-3)}$, with sign same as that of $\mu_3$.

$$y_0 = \frac{N\gamma^{p-1}}{\Gamma(p-1)},$$

$$sk. = \frac{2\sqrt{p-3}}{p}.$$

$$\text{Origin} = \text{mean} - \frac{\gamma}{p-2}.$$

$$\text{Mode} = \text{mean} - \frac{2\gamma}{p(p-2)}.$$

**Type VI.** The equation is

$$y = y_0 (x-a)^{q_2} x^{-q_1}.$$

The formulas are:

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{6 + 3\beta_1 - 2\beta_2},$$

$$e = \frac{4r^2}{16(r+1) + \beta_1(r+2)^2},$$

$$1 - q_1 = -\frac{r}{2} + \frac{r+2}{4}\sqrt{e\beta_1},$$

$$1 + q_2 = -\frac{r}{2} - \frac{r+2}{4}\sqrt{e\beta_1},$$

$$a = \frac{r\sigma}{\sqrt{e}}.$$

$$y_0 = \frac{N a^{q_1 - q_2 - 1}\,\Gamma(q_1)}{\Gamma(q_1 - q_2 - 1)\Gamma(q_2 + 1)}.$$

$$\text{Origin} = \text{mean} - \frac{a(q_1 - 1)}{q_1 - q_2 - 2}.$$

$$\text{Mode} = \text{mean} - \frac{1}{2}\frac{\mu_2}{\mu_3}\cdot\frac{r+2}{r-2}$$

**Type VII.** The equation is:

$$y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m}.$$

The formulas are:

$$m = \frac{5\beta_2 - 9}{2(\beta_2 - 3)},$$

$$a^2 = \frac{2\mu_2\beta_2}{\beta_2 - 3}.$$

$$y_0 = \frac{N}{a\sqrt{\pi}}\,\frac{\Gamma m}{\Gamma(m - \tfrac{1}{2})}$$

**Normal Curve.** The equation, as was proved in Chapter VI, is

$$y = \frac{N}{\sqrt{2\pi\sigma}}\cdot e^{-\frac{x^2}{2\sigma^2}},$$

and the curve was discussed in that chapter.

## APPENDIX II


BLAKEMAN, J.

"On the Tests for Linearity of Regression in Frequency Distributions", *Biometrika, Vol. IV, pp. 332 et seq.*

BLAKEMAN, J. and PEARSON, K.

"On the Probable Error of Mean Square Contingency", *Biometrika, Vol. V, pp. 191 et seq.*

BOWLEY, A. L.

"Measurements of Groups and Series", *C. and E. Layton, 1903.*
"Relation between the Accuracy of an Average and that of its Constituent Parts", *Jour. Roy. Stat. Soc., Dec. 1897.*
"The Measurement of the Accuracy of an Average", *Jour. Roy. Stat. Soc., Dec., 1911.*

BRAVAIS, A.

"Analyse matematique sur les probabilités des erreurs de situation d'un point", *Memoires présentés par divers savants a L'academie Royale des Sciences de L'institute de France, sciences matematique et physique, IIe série, t. IX, 1846, p. 255.*

BROWN, GREENWOOD and WOOD.

"A Study of Index Correlation", *Jour. Roy. Stat. Soc., Feb. 1914.*

CAVE, BEATRICE and PEARSON, K.

"Numerical Illustrations of the Variate-Difference Correlation Method", *Biometrika, Vol. X, pp. 340 et seq.*

EDGEWORTH, F. Y.

"On the Method of Least Squares", *Phil. Mag., Vol. XVI, Ser. 5, 1883, pp..360 et seq.*
"On Theory of Errors of Observation and the First Principles of Statistics", *Camb. Phil. Trans., Vol. XIV, pp. 138 et seq.*
"Problems in Probability", *Phil. Mag., Vol. XXII, Ser. 1, 1886, pp. 374 et seq.*
"On a New Method of reducing Observations relating to several Quantities", *Phil. Mag., Vol. XXIV, Ser. 5, 1887, pp. 222 et seq.* and *Vol. XXV, 1888, pp. 184 et seq.*
"On Correlated Averages", *Phil. Mag., Vol. XXXIV, Ser. 5, 1892, pp. 190 et seq.*

EDGEWORTH, F. Y.

"The Asymmetrical Probability Curve", *Phil. Mag., Vol. XLI, 1896, pp. 90 et seq.*

"Representation of Statistics by Mathematical Formulas", *Jour. Roy. Stat. Soc., Dec. 1898; Sept. 1899; June 1899; Mar. 1899; Mar. 1900.*

"The Law of Error", *Camb. Phil. Trans., Vol. XX, 1905; pp. 36-65 and 113-141.*

"On the Generalized Law of Error of Great Numbers", *Jour. Roy. Stat. Soc., Sept. 1906.*

"On the Representation of Statistical Frequencies by a Series", *Jour. Roy. Stat. Soc., Mar. 1907.*

"On the Representation of Statistics by Analytical Geometry", *Jour. Roy. Stat. Soc., 1914; Feb. pp. 300-312; Mar. 415-432; May, 653-671; June, 724-749; July. 838-852.*

Article on "Probability" in the Encyclopedia Brittanica, *Eleventh Edition.*

EDGEWORTH, F. Y. and BOWLEY, A. L.

"Methods of Representing Statistics of Wages and other Groups not Fulfilling the Normal Law of Error", *Jour. Roy. Stat. Soc., June, 1902.*

ELDERTON, W. PALIN.

"Frequency Curves and Correlation", *C. and E. Layton, London, 1906.*

ELLIS, LESLIE.

"The Method of Least Squares", *Camb. Phil. Trans., Vol. VIII, pp. 1 et seq.*

FISHER, R. A.

"On an Absolute Criterion for Fitting Frequency Curves", *Messenger, Vol. XLI, pp. 165-160.*

GALTON, FRANCIS.

"Family Likeness in Stature", *Proc. Roy. Soc., Vol. XL, 1886; pp. 42 et seq.*

"Family Likeness in Eye-Color", *Proc. Roy. Soc., 1886; Vol. XL, pp. 402 et seq.*

GALTON, FRANCIS.

"Correlations and their Measurement", *Proc. Roy. Soc., Vol. XLV, 1888, pp. 135 et seq.*

"The most Suitable Proportions between First and Second Prizes", *Biometrika, Vol. I, pp. 385 et seq.*

HERON, DAVID.

"On the Probable Error of a Partial Coefficient", *Biometrika, Vol. VII, pp. 411 et seq.*

"The Danger of Certain Formulae Suggested as Substitutes for the Correlation Coefficient", *Biometrika, Vol. VIII, pp. 109 et seq.*

HOOKER, R. H.

"Correlation of the Marriage Rate with Trade", *Jour. Roy. Stat. Soc., Sept. 1901.*

"Correlation of the Weather and Crops", *Jour. Roy. Stat. Soc., Mar. 1907.*

ISSERLIS, L.

"On the Partial Correlation Ratio", *Biometrika, Vol. X, pp. 391 et seq.,* also *Vol. XI.*

"The Application of Solid Hypergeometrical Series to Frequency Distribution in Space", *Phil. Mag., Vol. XXVIII, Ser. 6, 1914, pp. 379 et seq.*

KEYNES, J. M.

"Principal Averages and the Laws of Error which lead to them", *Jour. Roy. Stat. Soc., Feb. 1911.*

NIXON, J. W.

"An Experimental Test of the Normal Law of Error", *Jour. Roy. Stat. Soc., June, 1913.*

PEARSON, KARL.

"Mathematical Contributions to the Theory of Evolution",

I. "On the Dissection of Asymmetrical Frequency Curves", *Phil. Trans., 1894, Vol. CLXXXV, A, part I, pp. 187 et seq.*

II. "Skew Variations in Homogeneous Material", *Phil. Trans., 1895, Vol. CLXXXVI, A, pp. 343 et seq.*

III. "Regression, Heredity and Panmixia", *Phil. Trans., 1896, Vol. CLXXXVII A, pp. 253 et seq.*

IV. "On Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation", *Phil. Trans. 1898, Vol. CXCI A, pp, 229 et seq. (In Collaboration with L. N. G. Filon.)*

V. "On the Reconstruction of the Stature of Prehistoric Races", *Phil. Trans., 1892, Vol. CXCII A, pp. 169 et seq.*

VI. "General Selection", *Phil. Trans., 1899, Vol. CXCII A, pp. 257 et seq.*

VII. "On the Correlation of Characteristics not Quantitatively Measurable", *Phil. Trans., 1909, Vol. CXCV A, pp. 1 et seq.*

VIII. "On the Inheritance of Characters not Capable of Exact Quantitative Measurements", *Phil. Trans., 1901, Vol. CXCV A, pp. 79 et seq.*

IX. "On the Principles of Homotyposis and its Relation to Heredity, to the Variability of the Individual and to that of the Race", *Phil. Trans., 1901, Vol. CXCVII A, pp. 285 et seq.*

PEARSON, KARL.

"Mathematical Contributions to the Theory of Evolution" — Continued.

X. "Supplement to a Memoir on Skew Variation", *Phil. Trans., 1901, Vol. CXCVII A, pp. 445 et seq.*

XI. "On the Influence of Natural Selection on the Variability and Correlation of Organs", *Phil. Trans., Vol. CC, A, 1903, pp. 1 et seq.*

XII. "On a Generalized Theory of Alternative Inheritance with Special Reference to Mendel's Law", *Phil. Trans., 1904, Vol. CCIII A, pp. 53 et seq.*

XIII. "On the Theory of Contingency and its Relation to Association and Normal Correlation", *Drapers' Co. Res. Mem., Biometric Series I, Dulau & Co., London, 1904.*

XIV. "On the General Theory of Skew Correlation and Non-Linear Regression", *Drapers' Co. Res. Mem., Dulau & Co., 1905.*

XV. "A Mathematical Theory of Random Migration", *Drapers' Co. Res. Mem., Biometric Series II, 1906.* (*In Collaboration with Blakeman*).

XVI. "On Further Methods of Determining Correlation", *Drapers' Co. Res. Mem., Biometric Series IV, Dulau & Co., London, 1907.*

XVIII. "On a Novel Method of Regarding Association, etc.", *Biometric Series VII, 1912, Drapers' Co. Res. Mem.*

"On a Form of Spurious Correlation due to Indices", *Proc. Roy. Soc., Vol. LX, 1897, pp. 489 et seq.*

"On a Criterion that a given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling", *Phil. Mag., Ser. 5, Vol. L, 1900, pp. 157 et seq.*

"On Lines and Planes of Closest Fit to Systems of Points in Space", *Phil. Mag., Ser. 6, Vol. II, 1901, pp. 559 et seq.*

"On the Systematic Fitting of Curves to Observations and Measurements", *Biometrika, I, pp. 265 et seq. and Biometrika, II, pp. 1 et seq.*

"On the Probable Errors of Frequency Constants", *Biometrika, II, pp. 273 et seq.; also Vol. IX, pp. 1 et seq.*

"Elementary Proof of Sheppard's Formulae, etc.", *Biometrika, Vol. III, pp. 308 et seq.*

"On the Generalized Probable Error in Multiple Normal Correlation", *Biometrika, Vol. VI, 1908, pp. 59 et seq.* With *Alice Lee*

"On a New Method of Determining Correlation between a Measured Character A and a Character B of which only the Percentage of Cases wherein B exceeds (or falls short of) a given Intensity is recorded for each Grade of A", *Biometrika, Vol. VII, 1909, pp. 96 et seq.*

PEARSON, KARL.

"Mathematical Contributions to the Theory of Evolution"—Concluded.

"On a New Method of Determining Correlation when one Variable is given by Alternative and the other by Multiple Categories", *Biometrika, Vol. VII, 1910, pp, 248 et seq.*

"On a Correction to be made to the Correlation ratio $\eta$", *Biometrika, Vol. VIII, pp. 254 et seq.*

'On the Probable Error of a Coefficient of Correlation as found from a fourfold Table", *Biometrika, Vol. IX, pp. 22, et seq.*

"On the Measurement of the Influence of 'Broad Categories' on Correlation", *Biometrika, Vol. IX, pp. 166, et seq.*

PEARSON, K. (Editor).

"Tables for Statisticians and Biometricians", *Cambridge University Press, 1914.*

PEARSON, K. and HERON, DAVID.

"On Theories of Association", *Biometrika, Vol. IX, pp. 158 et seq.*

PERSONS, WARREN.

"The Correlation of Economic Statistics", *Amer. Stat. Assoc., Vol. XII, Dec. 1910.*

SHEPPARD, W. F.

"On Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation", *Phil. Trans., 1899, Vol. CXCII, A, pp. 101 et seq.*

"On the Calculation of the most probable Values of the Frequency Constants for Data arranged according to equi-distant Divisions of a Scale", *Proc. Lon. Math. Soc., Vol. XXIX, pp. 353-380.*

"On the Use of Auxiliary Curves in Statistics of Continuous Variates", *Jour. Roy. Stat. Soc., Sept. 1900.*

SNOW, E. C.

"The Application of the Method of Multiple Correlation to the Estimate of Post Censal Population", *Jour. Roy. Stat. Soc., May, 1911.*

SPEARMAN, C.

"The Proof and Measurement of Association between Two Things.", *Amer. Jour. of Psychology, Vol. XV, 1904, pp. 88 et seq.*

"Demonstration of Formulae for True Measurement of Correlation", *Amer. Jour. of Psych., Vol. XVIII, 1907, pp. 161 et seq.*

"A Foot-rule for Measuring Correlation", *Brit. Jour. of Psych., Vol. II, 1906 pp.* RT *et seq.;* also *Vol. II, part* V, *pp. 107-108.*

"Correlation calculated from Faulty Data", *Brit. Jour. of Psych., Vol. III, 1910, pp. 271 et seq.*

"STUDENT".

"The Elimination of Spurious Correlation due to Position in Time or Space", *Biometrika, Vol. X, pp. 799 et seq.*

YULE, G. U.

"On the Significance of Bravais' Formulae for Regression, etc., in the case of Skew Correlation", *Proc. Roy. Soc., Vol. LX, 1897, pp. 477 et seq.*

"On the Association of Attributes in Statistics", *Phil. Trans., 1900, Vol. CXCIV, A, pp. 257 et seq.*

"On the Theory of Consistence of Logical Class Frequencies and its Geometrical Representations", *Phil. Trans., 1901, Vol. CXCVII, A, pp. 91 et seq.*

"On the Theory of Correlation for any Number of Variables treated by a New System of Notation", *Proc. Roy. Soc., Ser. A, Vol. LXXIX, 1907, pp. 182 et seq.*

"The Application of the Methods of Correlation to Social Economic Statistics", *Jour. Roy. Stat. Soc., Dec. 1909.*

"On Interpretation of Correlation between Indices or Ratios", *Jour. Roy. Stat. Soc., June, 1910.*

"On the Methods of Measuring Association between Two Attributes", *Jour. Roy. Stat. Soc., May, 1912.*